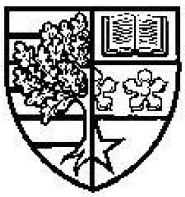
Heriot-Watt University



Efficient Numerical Schemes for Porous Media Flow

Antoine Tambue

October 19, 2010

Submitted for the degree of
Doctor of Philosophy in Mathematics
on completion of research in the
Department of Mathematics,
School of Mathematical and Computing Sciences.

Declaration

I hereby declare that the work presented in this thesis was carried out by myself at Heriot-Watt University, except where due acknowledgement is made, and not been submitted for any other degree.

Signature of Antoine Tambue (Candidate)

Signature of Supervisors: Professor Gabriel Lord Dr Sebastian Geiger

ABSTRACT

Partial differential equations (PDEs) are important tools in modeling complex phenomena, and they arise in many physics and engineering applications. Due to the uncertainty in the input data, stochastic partial differential equations (SPDEs) have become popular as a modelling tool in the last century. As the exact solutions are unknown, developing efficient numerical methods for simulating PDEs and SPDEs is a very important while challenging research topic. In this thesis we develop efficient numerical schemes for deterministic and stochastic porous media flows. More schemes are based on the computing of the matrix exponential functions of the non diagonal matrices, we use new efficient techniques: the real fast Léja points and the Krylov subspace techniques.

For the deterministic flow and transport problem, we consider two deterministic exponential integrator schemes: the exponential time differential stepping of order one (ETD1) and the exponential Euler midpoint (EEM) with finite volume method for discretization in space. We give the time and space convergence proof for the ETD1 scheme and illustrate with simulations in two and three dimensions that the exponential integrators are efficient and accurate for advection dominated deterministic transport flow in heterogeneous anisotropic porous media compared to standard semi implicit and implicit schemes.

For the stochastic flow and transport problem, we consider the general parabolic SPDEs in a Hilbert space, using the finite element method for discretization in space (although finite difference or finite volume can be used as well). We use a linear functional of the noise and the standard Brownian increments to develop and give convergence proofs of three new efficient and accurate schemes for additive noise, one called the modified semi–implicit Euler-Maruyama scheme and two stochastic exponential integrator schemes, and two stochastic exponential integrator schemes for multiplicative and additive noise. The schemes are applied to two dimensional flow and transport.

Dedication

To my late father, Kamga Joseph and my mother Ngamgo Madeleine.

Acknowledgements

I would like to thank the almighty God, who created the whole universe and who makes all things possible.

I am deeply indebted to my advisors, Prof. Gabriel Lord (Department of Mathematics) and Dr. Sebastian Geiger (Institute of Petroleum Engineering), for their guidance and encouragement during my study at Edinburgh. During our numerous research discussions and frequent personal contact, they keep inspiring me both as experienced scientists and good friends. They have always wanted the best for me, and supported every endeavor I have made. I consider myself very lucky to have had the opportunity to work with the two open minded people with complementary backgrounds.

I would like to express my deepest gratitude to Dr. Bernd Schroers who encouraged me to pursue my Ph D study here at Heriot Watt University and Prof Niel Turok for his foresight in founding the African Institute for Mathematical Sciences (AIMS), my former institute, where I met Prof. Gabriel Lord and Dr. Bernd Schroers as visiting lecturers.

I wish to thank all my lecturers at the University of Dschang and University of Yaounde I in Cameroon for their training and continuous support. Especially Dr. Njifenjou Abdou, Dr. Kengne Emmanuel, Prof. Marcel Dossa, Prof. Tayou Simo Jacques, Prof. Norbert Noutchegueme, Prof Nguetsing Gabriel, Prof. Bekole David and Prof. Wamon Francois.

I must also thank ORSAS and Heriot Watt University for funding my PhD research work.

I would also like to thank all my colleagues for the nice times we had together during the work and outside the University, especially the "AIMS family" (Djeundje Biatat Viani, Issa Karambal, Gamado Kokouvi, Iyabo Ann Adamu, Dr. Nneoma Ogbonna, Christine Ayawoa Dagbovie), Jennifer Reynolds and Yusoff Yumn.

I would like to thank all my previous classmate and friends. Especially Alex Chime,

Fotso Tamko Richard, Dr. Feunou Kamkui Bruno, Takam Takougang Eric–Martial, Elisabeth Teudjeu Kemayou, Laurent Tchoualag, Benoit Sehba, Franck Kalala, Tagne Serges Bertrand (of late), Tuekam Simo Blaise.

Finally, I would like to thank my mother Ngamgo Madeleine, my family (Tambue Clementine, Tambue Kamga Serges Lions, Tambue Ngamgo Gaelle), my grand mother (of late), my brothers, my sisters, my uncles, my aunts and my grand father's family (Wabo Tagatsi's family) for their constant love and support.

Contents

List of Tables

List of Figures

Introduction

Many practical problems arising in the real life applications can be modeled by time dependent partial differential equations (PDEs). Advection and diffusion can transport chemically reactive components such as dissolved minerals, colloids, or contaminants, over long distances through the highly heterogeneous porous media comprising geological formations. It is hence a fundamental process in many geo-engineering applications, including oil and gas recovery from hydrocarbon reservoirs, groundwater contamination and sustainable use of groundwater resources, storing greenhouse gases (e.g, CO_2) or radioactive waste in the subsurface, or mining heat from geothermal reservoirs. One of the fundamental challenges is to forecast these processes accurately because the permeability in heterogeneous porous and fractured media typically varies over orders of magnitude in space and possibly time (e.g, [1, 2]). This causes highly variable flow fields where local transport can be dominated entirely by either advection (Péclet number larger than one) or diffusion (Péclet number less than one), leading to macroscopic mixing and "anomalous transport" that is characterised by early breakthrough of solutes or contaminants and long tailing at late time [3]. Chemical reaction rates and equilibrium constants can vary in a similar manner, giving rise to complex mixing-induced reaction patterns at the macro-scale because chemical reactions rates can dominate locally over transport rates or vice versa (e.g, [4–6]).

Predicting the spatial spreading and mixing of reactive solutes in field applications hence requires the efficient and accurate numerical solution of advection-diffusion-reaction equations (ADR) which resolve the wide range in flow velocities and reaction rates. This is particularly important because the exact spatial distribution of the permeability field and reaction rates is commonly unknown and therefore a large number of simulations must be run to quantify the uncertainty of the transport behaviour [7], for example to forecast the possible arrival of highly toxic contaminants at a groundwater well and design adequate

remediation schemes.

The efficient time integration of evolution equations requires particular methods. For quite a long time, implicit and linearly implicit methods were the methods of choice. These methods, however, need at each time step a solution of large systems of nonlinear equations. This can be the bottleneck in computations. In recent years, exponential integrators have become an attractive alternative in many situations. In contrast to classical methods, they do not require the solution of large linear systems. Instead they make explicit use of the matrix exponential and related matrix functions. The family of exponential integrators date back to the 1960's (see [8,9]). These methods are based on approximating the corresponding integral formulation of the non-linear part of the differential equation, solving the linear part exactly, and computing the exponential of a matrix, a notorious problem in numerical analysis [10]. However, new developments in computing exponential matrix functions has revived interest for these methods.

Due to the lack of information on the input data, Stochastic Partial Differential Equations (SPDEs) became popular since the last century. Like mathematical models using PDEs, many SPDEs models do not admit analytical solutions and we must look for accurate and efficient numerical schemes to approximate them.

This thesis presents our contribution to numerical schemes for flow and transport problem in porous media. Chapter 1 to Chapter 3 focus on deterministic flow and transport while Chapter 4 to Chapter 7 focus on a general stochastic parabolic partial differential equations where stochastic flow and transport problem is a particular case. Our special contributions are in Chapter 3 and Chapter 5 to Chapter 7.

Chapter 1 presents the fundamental physics of single–phase flow and transport in saturated porous media. The main aim is to recall the basic of equations and constants used in this thesis. These equations are described by Darcy's law and mass balance equations.

In Chapter 2, we give a rigorous statement of the model problems using some fundamental notions from functional analysis, and the corresponding classical finite volume space discretization. In this chapter, we review results from Sobolev space [11] and semi group theory [12].

Chapter 3 presents the exponential integrator time stepping schemes for advection–diffusion–reaction (ADR). In this chapter, we investigate two exponential time integrators,

the second-order accurate Exponential Euler Midpoint (EEM) scheme [13] and Exponential Time Differencing of order one (ETD1) [14] for advection-dominated reactive transport in anisotropic and heterogeneous porous media. The time and space convergence is given for ETD1 scheme. The exponential matrix function, the so called φ_1- function is computed with the real fast Léja points and Krylov subspace techniques. All our numerical examples, which include advection–diffusion–reaction simulations performed on the classical SPE10 test case [1], demonstrate that exponential integrators are highly competitive compared to standard semi-implicit and implicit time integrators. Chapter 3 extends work published in our papers [15, 16].

We start stochastic analysis from Chapter 4, using the finite element method [11, 17] for space discretizations. We note that extension to finite differences or finite volumes methods would be possible. In Chapter 4, we give a rigorous statement of nonlinear SPDEs and corresponding standard numerical stochastic schemes. In Chapter 5, we introduce a new scheme for SPDEs driven by additive space-time noise using a linear functional of the noise with a semi–implicit Euler–Maruyama method to discretize in time. We give convergence proofs in the root mean square L^2 norm for a diffusion reaction equation and in root mean square H^1 norm in the presence of advection under some regularity of the noise. We present numerical results for a linear reaction diffusion equation in two dimensions as well as a nonlinear example of a two-dimensional stochastic advection–diffusion–reaction equation. The analysis and numerics shows that we have better convergence properties than for the standard semi–implicit Euler–Maruyama method. Chapter 5 extends the work presented in our paper [18].

Chapter 6 introduces two new schemes for SPDEs driven by additive space-time noise. We upgraded two deterministic exponential schemes to stochastic exponential schemes using a linear functional as in Chapter 5. As in the deterministic schemes, we compute the exponential matrix functions by using the real fast Léja points and Krylov subspace techniques. We consider noise that is white in time and either in H^1 or H^2 in space and give convergence proofs in the mean square L^2 norm for a diffusion reaction equation and in mean square H^1 norm in the presence of an advection term. We present results for a linear reaction diffusion equation in two dimensions as well as a nonlinear example of two-dimensional stochastic advection–diffusion–reaction equation motivated from

realistic porous media flow, which shows better convergence properties over the standard semi–implicit Euler–Maruyama method and the new modified semi–implicit scheme built in Chapter 5. This chapter is presented in our paper [19].

Chapter 7 extends the deterministic ETD1 and exponential Lawson schemes to stochastic exponential integrators for more general noise (additive and multiplicative space time noise). The time step discretization for the noise used here is the standard Brownian increment. We give the convergence proofs in the root mean square L^2 norm. Again, we compute the exponential matrix functions using the real fast Léja points and the Krylov subspace techniques. We present results for a linear reaction diffusion equation in two dimensions as well as a nonlinear example of two-dimensional stochastic advection diffusion reaction equation motivated from realistic porous media flow.

Chapter 1

Flow and transport in porous media: Basic notions

In the first chapter of this thesis, we present the basic notions for single–phase fluid flow and transport in saturated porous media. The main aim is to recall the key equations and constants used in this thesis. These equations represent momentum and mass balance in porous media and are given by Darcy's law and two conservation equations respectively. The conservation equation of fluid flow, assuming incompressibility of the fluid, is given by the divergence equation. The conservation equation describing the transport of a dissolved and chemically reactive species contains and advection, dispersion (includng diffusion) and reaction term. Examples for reactions are adsorption, or radioactive decay.

1.1 Definition of basis concepts in porous media

More information about the content of this chapter can be found in the standard texts [20, 21].

1.1.1 Porous medium

A porous medium is a medium which contains void space called pores allowing fluid to flow through the medium. Many natural substances such as rocks, soils, biological tissue (e.g. bones), and man-made materials such as cements, foams and ceramics can be considered as porous media. A porous medium is characterised by its porosity, permeability as well as the

properties of its constituents (solid matrix and fluid). In this thesis our main porous medium will be the geological reservoir (a porous medium which contains different types of rocks). A geological reservoir is complex as the permeabilty and porosity are highly uncertain functions of space and possibly time. If the reservoir contains oil or gas, it is called an oil reservoir or hydrocarbon reservoir, if it contains water it is called a groundwater reservoir and if it contains heat it is called a geothermal reservoir.

1.1.2 Porosity

The porosity ϕ of a porous medium is a fraction of pore space volume, i.e.

$$\phi = \frac{\textbf{Pore space volume}}{\textbf{Total volume of the porous medium}}. \tag{1.1}$$

In all this work ϕ will be used for effective porosity i.e. the fraction of connected pores where fluids can actually flow in. Figure 1.1 shows a statistically generated porosity field of a potential oil reservoir [1] while Table 1.1 gives typical values of the porosity of some rocks.

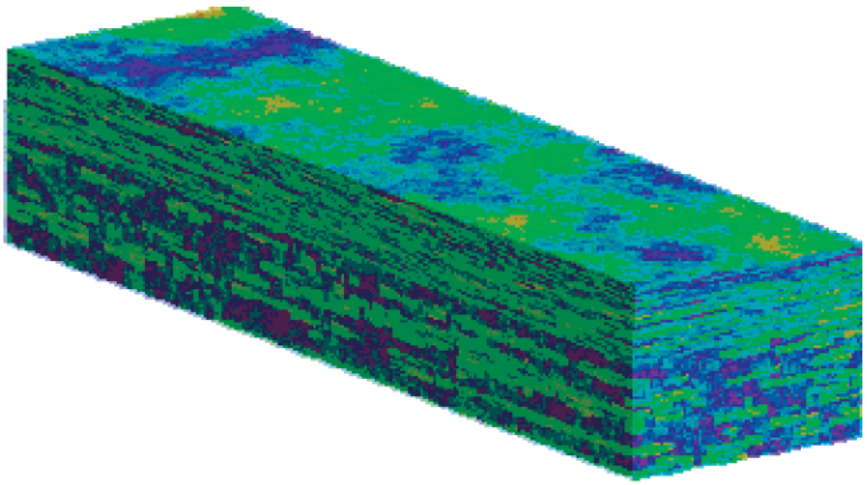


Figure 1.1: A statistically generated porosity field of a potential oil reservoir (the SPE 10 model). Blue values represent low porosities and red values high porosities.

1.1.3 Permeability

The permeability (or intrinsic permeability) $\mathbf{k}$ with units of [m^2] or [Darcy] is the quantity which determines the ease at which fluids can flow through a porous medium, i.e. the

Table 1.1: Average of the porosity of some rocks. Taken from [20].

Material	Porosity (%)	Material	Porosity (%)
Gravel, Coarse	28	Loess	49
Gravel, mesium	32	Peat	92
Gravel, fine	34	Schist	38
Sand, coarse	39	Siltstone	35
Sand, medium	39	Claystone	43
Sand, fine	43	Shale	6
Silt	46	Till, predominantly silt	34
Clay	42	Till, predominantly sand	31
Sandstone, fine grained	33	Tuff	41
Sandstone, medium grained	37	Basalt	17
Limestone	30	Gabbo, weathered	43
Dolomite	26	Granite, weathered	45

resistance to fluid flow. In a complex porous medium the permeability is usually a full tensor given by

$$\mathbf{k} = \begin{pmatrix} k_{xx} & k_{xy} & k_{xz} \\ k_{yx} & k_{yy} & k_{yz} \\ k_{zx} & k_{zy} & k_{zz} \end{pmatrix}. \tag{1.2}$$

A porous medium is anisotropic with respect to the permeability if it is directionally dependent and isotropic if not. The permeability in an isotropic porous medium is given by $\mathbf{k} = k\mathbf{I}_3$ where $\mathbf{I}_3$ is the 3×3 identity tensor. A key challenge is that the distribution of the permeability is commonly highly uncertain in subsurface reservoirs while the permeability values vary over several orders of magnitude in space.

In hydrogeology, the permeability combined with the fluid properties viscosity [Pa s] and density [kg m^{-3}], as well as the constant of gravity [m s^{-2}], to form the hydraulic conductivity $\mathbf{K}$ [m s^{-1}] given by

$$\mathbf{K} = \frac{\rho g}{\mu} \cdot \mathbf{k}. \tag{1.3}$$

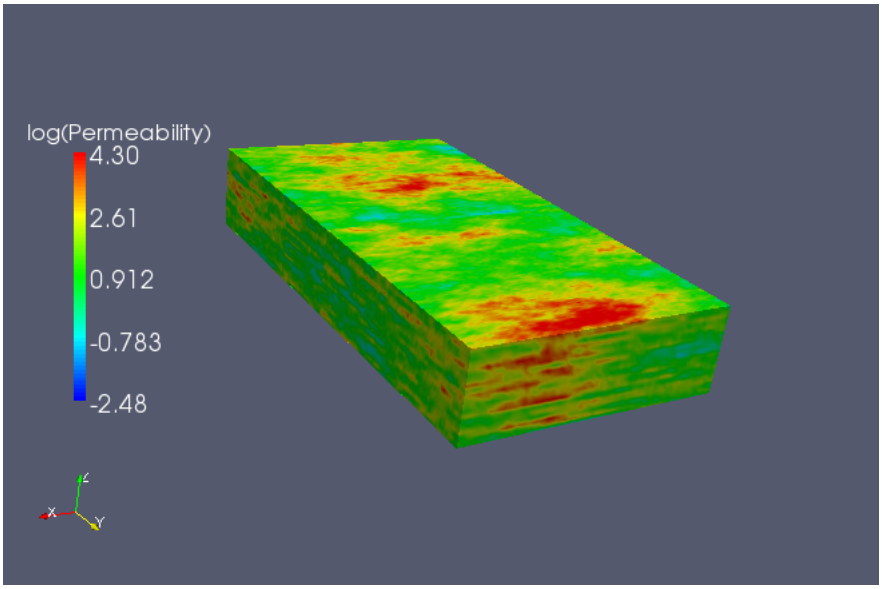


Figure 1.2: Statistically generated permeability field ($y-$ direction) of the first upper 20 layers of the SPE 10 case. Note the $\log_{10}$ scale and that permeability varies over 6 orders of magnitude.

Table 1.2 illustrates the range of the hydraulic conductivities (assuming pure water at room conditions) for some geological media.

1.2 Darcy's Law

In 1856, Henry Darcy investigated the flow of water in a vertical, saturated, homogeneous sand filter (see Figure 1.3), in connection with the Dijon city's fountains. From his experiments, varying the length and diameter of the column, the porous material in it, and the water levels in inlet and outlet reservoirs, he concluded that the rate of flow Q (volume of water passing per unit time) through a sand column of length L and constant cross-sectional area A is:

- Proportional to the cross-sectional area A of the column.
- Proportional to the difference in water level elevations h_3 and h_4 in the inflow and outflow reservoirs of the column respectively and inversely proportional to the column length L.

Table 1.2: Hydraulic conductivity of the water in different porous media. Taken from [20].

Rocks	Hydraulic conductivity (m/sec)
Unconsolidated sedimentary deposits	
Gravel	3×10^{-2} to 3×10^{-4}
Coarse sand	6×10^{-3} to 9×10^{-7}
Medium sand	6×10^{-4} to 9×10^{-7}
Fine sand	2×10^{-4} to 2×10^{-7}
Silt, loess	2×10^{-5} to 1×10^{-9}
Till	2×10^{-6} to 1×10^{-12}
Clay	5×10^{-9} to 1×10^{-11}
Unweathered marine clay	2×10^{-9} to 8×10^{-13}
Sedimentary rocks	
Karst limestone	2×10^{-2} to 1×10^{-6}
Limestone and dolomite	6×10^{-6} to 1×10^{-9}
Sandstone	6×10^{-6} to 3×10^{-10}
Shale	2×10^{-9} to 1×10^{-13}
Crystalline rocks	
Permeable basalt	2×10^{-2} to 4×10^{-7}
Fractured igneous and metamorphic	3×10^{-4} to 8×10^{-9}
Basalt	4×10^{-7} to 2×10^{-11}
Unfractured igneous and metamorphic	2×10^{-10} to 3×10^{-14}
Weathered grantie	5×10^{-5} to 3×10^{-6}

When combined, these conclusions give the famous Darcy's formula, or Darcy's law

$$Q = \mathbf{K} \cdot A \frac{h_3 - h_4}{L} = \mathbf{K} \cdot A \frac{\Delta h}{L}. \tag{1.4}$$

See [22] for more details.

It is important to mention that h is defined as the piezometric head, given in [22] by

$$h = z + \frac{p}{\rho g}, \tag{1.5}$$

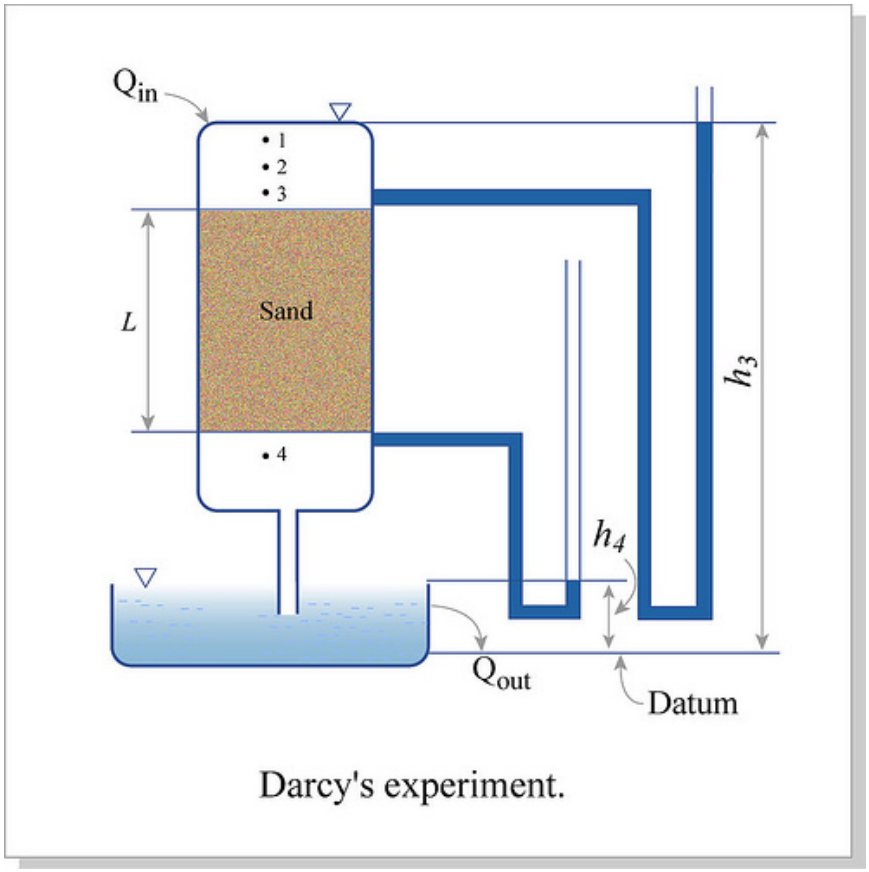


Figure 1.3: Darcy's experiment. Adapted from [23].

where z is the elevation and p the pressure in the fluid at the considered point, ρ and g are defined in the previous section. When L tends toward 0, Darcy's law (1.4) allows to have the following Darcy's velocity denoted by $\mathbf{q}$

$$\mathbf{q} = \frac{Q}{A} = -\mathbf{K}\nabla h = -\frac{\mathbf{k}}{\mu} \cdot (\nabla p + \rho g \nabla z) = -\frac{\mathbf{k}}{\mu} \cdot (\nabla p + \rho \mathbf{g}) \,. \tag{1.6}$$

The Darcy's law velocity $\mathbf{q}$ is also called a specific discharge and is related to the fluid's velocity $\mathbf{v}$ and porosity ϕ by

$$\mathbf{v} = \frac{\mathbf{q}}{\phi}. \tag{1.7}$$

The equation (1.6) is called **equation of motion** of the fluid in the porous media. The equation of motion is equivalent to the Navier-Stokes equation in fluid dynamics and represents the momentum balance for fluid in porous media.

Given that in the equation of motion (1.6) we have two unknowns $\mathbf{q}$ (or $\mathbf{v}$) and h (or p), we need another equation to be able to solve the flow problem. This equation will come

from mass conservation (balance) of the fluid.

1.3 Mass conservation equation

Consider the unit volume of porous medium centered at the point (x, y, z) (Figure 1.4) called Representative Elemental Volume (REVo). The law of conservation of mass requires that the change in mass per time is equal to the mass flowing in (M_{in}) minus mass flowing out (M_{out}) of the unit volume plus source (M_{source}), i.e.

$$\frac{\partial M}{\partial t} = M_{in} - M_{out} + M_{source}. \tag{1.8}$$

Let us denote by $\mathbf{J}$ the total mass flux per unit of time and per area, and Q' an internally

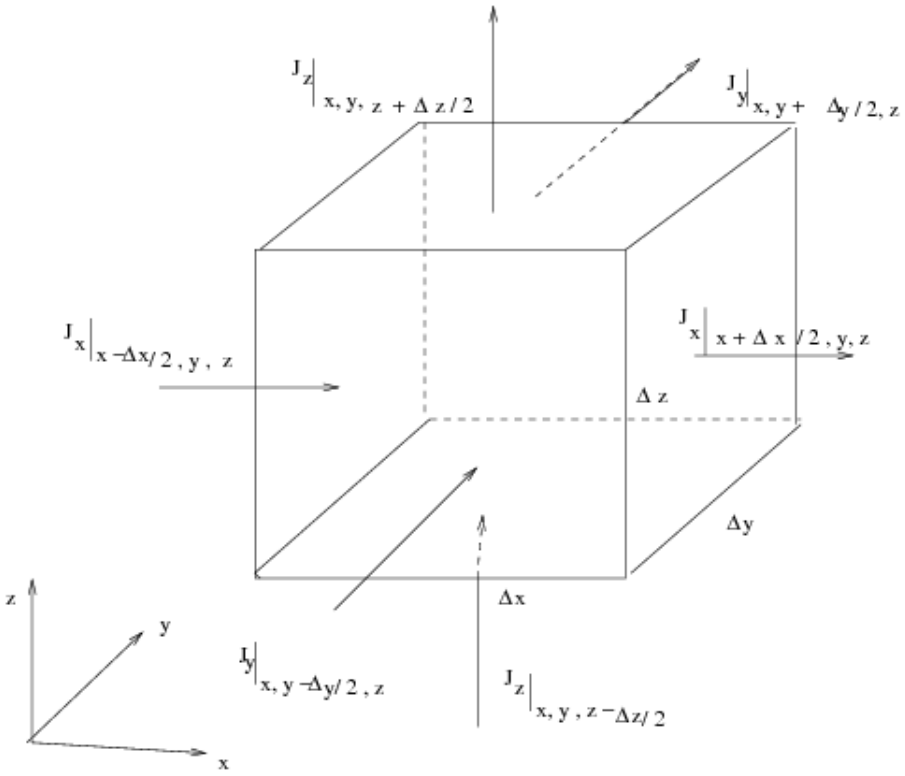


Figure 1.4: Representative Elemental Volume (REVo) of a medium with influx and outflux mass of a fluid at time t.

distributed source density [kg s^{-1} m^{-3}] which is positive if mass is produced and negative if mass is destroyed, We therefore have

$$\mathbf{J} = \rho\mathbf{q} = (\rho q_x, \rho q_y, \rho q_z)^T = (J_x, J_y, J_z)^T . \tag{1.9}$$

The change of mass of the fluid contained in the control volume represented by REVo is given by

$$\frac{\partial M}{\partial t} = \frac{\partial \phi\rho}{\partial t}\Delta x\Delta y\Delta z \tag{1.10}$$

and the internal source mass per unit of time by

$$Q'\Delta x\Delta y\Delta z. \tag{1.11}$$

The influx mass per unit of time is

$$J_x|_{(x-\Delta x/2,y,z)}\,\Delta y\Delta z + J_y|_{(x,y-\Delta y/2,z)}\Delta x\Delta z + J_z|_{(x,y,z-\Delta z/2)}\,\Delta x\Delta y \tag{1.12}$$

while the outflux mass per unit of time is

$$J_x|_{(x+\Delta x/2,y,z)}\Delta y\Delta z + J_y|_{x,y+\Delta y/2,z}\,\Delta x\Delta z + J_z|_{(x,y,z+\Delta z/2)}\,\Delta x\Delta y. \tag{1.13}$$

Applying the law of conservation of the mass yields

$$\begin{aligned}\Delta x\Delta y\Delta z\frac{\partial \phi\rho}{\partial t} &= \left(J_x|_{(x-\Delta x/2,y,z)} - J_x|_{(x+\Delta x/2,y,z)}\right)\Delta y\Delta z \\ &+ \left(J_x|_{(x,y-\Delta y/2,z} - J_x|_{(x,y+\Delta y/2,z)}\right)\Delta x\Delta z \\ &+ \left(J_x|_{(x,y,z-\Delta z/2)} - J_x|_{(x,y,z+\Delta z/2)}\right)\Delta x\Delta y + Q'\Delta x\Delta y\Delta z. \end{aligned}\tag{1.14}$$

Dividing (1.14) by $\Delta x\Delta y\Delta z$ and taking the limit as $\Delta x,\ \Delta y\,, \Delta z \to 0$ yields

$$\frac{\partial(\phi\rho)}{\partial t} = -\nabla\cdot(\rho\mathbf{q}) + Q'. \tag{1.15}$$

Indeed [24, 25]

$$\frac{\partial(\phi\rho)}{\partial t} = \frac{\partial\phi\rho}{\partial p}\,\frac{\partial p}{\partial t} = \frac{S_s}{g}\frac{\partial p}{\partial t}$$

where

$$S_s = g\ \frac{\partial\phi\rho}{\partial p}$$

is called specific storage (the volume of fluid that can be stored by compressing the porous medium and fluid itself) with unit $[m^{-1}]$.

Assume that the medium is not deformable, i.e. ϕ is independent of t, then using the relation $p = \rho g(h - z)$ yields

$$\begin{aligned}\frac{\partial p}{\partial t} &= g(h-z)\frac{\partial \rho}{\partial t} + \rho g \frac{\partial h}{\partial t} \\ &= \frac{S_s(h-z)}{\phi}\frac{\partial p}{\partial t} + \rho g \frac{\partial h}{\partial t},\end{aligned}$$

and

$$\frac{\partial p}{\partial t} = \left(\frac{1}{1 - \dfrac{S_s(h-z)}{\phi}} \right) \rho g \frac{\partial h}{\partial t}.$$

In most applications one has $\dfrac{S_s(h-z)}{\phi} \ll 1$ (see [24, 25]) and so we have the following approximation

$$\frac{\partial p}{\partial t} \approx \rho g \frac{\partial h}{\partial t}, \tag{1.16}$$

which yields

$$\frac{\partial(\phi\rho)}{\partial t} = \rho S_s \frac{\partial h}{\partial t}. \tag{1.17}$$

Using the relation (1.17) and the equation of motion (1.6) in the relation (1.15) yields the following mass conservation equation

$$S_s \rho \frac{\partial h}{\partial t} = \nabla \cdot (\rho\, \mathbf{K} \nabla h) + Q'. \tag{1.18}$$

The formulation of the mass conservation equation (1.18) in terms of pressure is given by

$$\frac{S_s}{g}\frac{\partial p}{\partial t} = \nabla \cdot \left(\frac{\rho}{\mu} \mathbf{k}\, (\, \nabla p + \rho \mathbf{g}) \right) + Q'. \tag{1.19}$$

If the spatial variation of ρ is negligible, (1.19) become

$$\frac{S_s}{\rho\, g}\frac{\partial p}{\partial t} = \nabla \cdot \left(\frac{\mathbf{k}}{\mu}\, (\nabla p + \rho \mathbf{g}) \right) + \frac{Q'}{\rho}. \tag{1.20}$$

The mass conservation equation (1.18) or (1.19) is a partial differential equation in h (or p) only since Q' and S_s are given. The value of h (or p) allows to compute $\mathbf{q}$ (or $\mathbf{v}$) directly from the equaton of motion (1.6) and the relation (1.7).

1.4 Flow and transport by advection, diffusion and chemical reaction

The aim of this section is to establish the conservation equation of a dissolved and chemical reactive component in porous media, which includes advection, diffusion and reaction. More details can be found in [20, 21].

Here we deal with a fluid (called the solution) which during motion, transports a dissolved substance (called solute). These substances could be toxic and possibly man-made such as contaminants transported in groundwater, or hydrocarbons transported in oil reservoirs, or CO_2 transported in saline aquifers. First of all we need to define different types of physical phenomena during flow and transport.

Advection is the movement of a solute along with the flowing fluid in porous media.

Let X denote the concentration of solute. The mass flux $\mathbf{J}_1$ due to advection is given by

$$\mathbf{J}_1 = \mathbf{q}X, \qquad X[\text{kg m}^{-3}], \;\; \mathbf{J}_1[\text{kg m}^{-2}\,\text{s}^{-1}]. \tag{1.21}$$

Diffusion is a molecular mass transport process in which a solute moves from areas of higher concentration to areas of lower concentration, driven by Brownian motion. This phenomena can occur in the absence of velocity (when solution is at rest). The mass flux $\mathbf{J}_2$ due to diffusion is given by Fick's law

$$\mathbf{J}_2 = -\mathbf{D}\nabla X, \tag{1.22}$$

where $\mathbf{D}$ is the diffusion tensor. In this thesis, in the case of anisotropic medium, we will assume that this tensor is relative to the principal directions of the anisotropic medium (in the case of anisotropic medium) and therefore takes the form

$$\mathbf{D} = \begin{pmatrix} D_x & 0 & 0 \\ 0 & D_y & 0 \\ 0 & 0 & D_z \end{pmatrix}. \tag{1.23}$$

It is important to notice that the negative sign before $\mathbf{D}$ in (1.22) indicates that the solutes moves towards the area of lower concentration.

In this work, the term **reaction** is used to indicate all chemical reactions such as biodegradation, sorption (adsorption and absorption), radioactive decay, fluid-rock reactions, etc. In our applications we will use the classical Langmuir isotherm to model the sorption of the transported species onto the rock surface, i.e.

$$R(X) = \frac{\lambda\beta X}{1+\lambda X}, \tag{1.24}$$

where the parameter λ is an adsorption constant and β the maximum amount of the solute that can be adsorbed.

Considering our control volume REVo in Figure 1.4, let us apply the concept of mass conservation to the solute. The total mass density by reaction depends on the type of reaction and will be denoted by a function $R_1(X)$. The total mass density of the source is denoted by Q_1. The change of mass of the solute contained in the control volume REVo is

$$\frac{\partial \phi X}{\partial t}\Delta x\Delta y\Delta z$$

and the total mass flux per unit of time is given by

$$\mathbf{J} = \mathbf{J}_1 + \mathbf{J}_2 = -\mathbf{D}\nabla X + \mathbf{q}X. \tag{1.25}$$

Let

$$R(\mathbf{x}, X) = R_1(\mathbf{x}, X) + Q_1(\mathbf{x}).$$

Applying the conservation of mass to the solute as in the previous section yields the following transport equation

$$\frac{\partial \phi X}{\partial t} = \nabla \cdot (\mathbf{D}\nabla X - \mathbf{q}X) + R(\mathbf{x}, X). \tag{1.26}$$

Remark 1.1 *Since equation (1.26) depends on the velocity* $\boldsymbol{q}$*, we need to solve equations (1.6) and (1.18) first to obtain a pressure (or head) field from the momentum equation and fluid velocity field from Darcy's law.*

In the case of uncertainties in the reaction term, which is almost always the case in porous media flow applications, the corresponding stochastic model is given [26] by

$$d(\phi X) = (\nabla \cdot (\mathbf{D}\nabla X - \mathbf{q}X) + R(\mathbf{x}, X))\, dt + b(X)dW, \tag{1.27}$$

where W is a space time noise introduced in Chapter 4 and $b(X)$ the noise intensity.

For illustration, assuming that the parameter β is uncertain in the classical Langmuir isotherm reaction function R given in (1.24), by setting

$$\beta = \beta_0 + \zeta,$$

where β_0 is a deterministic value and ζ a random forcing, the corresponding noise intensity in (1.27) is given by

$$b(X) = \frac{\lambda X}{1 + \lambda X}. \tag{1.28}$$

Chapter 2

The finite volume method for porous media flow and transport

In this chapter, we give a rigorous statement of the model problems using some fundamental notions from functional analysis, and the corresponding classical finite volume space discretization. We present the standard time stepping schemes usually used with some iterative linear solvers.

2.1 Well posedness of the system pressure–velocity

Let us start by presenting briefly the notation for the main function spaces and norms that we use in this thesis. We denote by $\|\cdot\|$ the norm associated to the inner product $(\cdot,\cdot)$ of the Hilbert space $H = L^2(\Omega)$ and by $|.|$ the standard Euclidian norm in $\mathbb{R}^n$, $n \in \mathbb{N}$. For a Banach space $\mathcal{V}$ we denote by $L(\mathcal{V})$ the set of bounded linear mapping from $\mathcal{V}$ to $\mathcal{V}$, $\|\cdot\|_{\mathcal{V}}$ the norm of $\mathcal{V}$ and $\|\cdot\|_{L(\mathcal{V})}$ the the norm of $L(\mathcal{V})$. More information about functional analysis and spaces can be found in [11, 12, 17]. We introduce further spaces below as required.

Let Ω be a bounded domain in $\mathbb{R}^d$, $d = 2, 3$. Consider the Darcy's velocity field and mass conservation law described in the first chapter by equations (1.6) and (1.19). Without loss of generality we assume here that $\mathbf{k} = k\mathbf{I}_d$.

Assuming that rock and fluids are incompressible, sources or sinks are absent and gravity

is negligible, the mass conservation law (1.19) become

$$\nabla \cdot \left(\frac{\rho \mathbf{k}}{\mu} \left(\nabla p + \rho \mathbf{g} \right) \right) = \nabla \cdot \left(\frac{\mathbf{k}}{\mu} \nabla p \right) = 0 \Leftrightarrow \nabla \cdot \mathbf{q} = 0, \tag{2.1}$$

which is known as divergence free velocity flow.

To solve the transport problem, we first find the two functions $\mathbf{q}$ and p such that

$$\begin{cases} \nabla \cdot \left(\dfrac{\mathbf{k}}{\mu} \nabla p \right) = 0 \\ q = -\dfrac{\mathbf{k}}{\mu} \nabla p \\ p = p_0 \qquad \text{in} \quad \partial\Omega_D^1 \\ \mathbf{q} \cdot \mathbf{n} = p_1 \quad \text{in} \quad \partial\Omega_N^1 \end{cases} \tag{2.2}$$

where $\partial\Omega = \partial\Omega_D^1 \cup \partial\Omega_N^1,\ \partial\Omega_D^1 \neq \emptyset$, $\mathbf{n}$ being the unit outward normal to $\partial\Omega_N^1$.
In pratice, the "no–flow" ($p_1 = 0$) condition is mostly used . Due to lack of data, it is almost impossible to obtain a detailed description of the distribution of the permeability $\mathbf{k}$ in the subsurface. To model the uncertainty in the permeability field, people usually introduce a certain level of random variability of $\mathbf{k}$ and assume that $\mathbf{k}$ is a stochastic field. As a result, the pressure p and the Darcy's velocity $\mathbf{q}$ are also the random fields. This will then show how the uncertainty in the input will propagate and affect the output of the model. Since $\mathbf{k}$ is modeled as a random field instead of the pressure–velocity system (2.2), we consider the problem of finding two random fields $q(\mathbf{x}, w)$ and $p(\mathbf{x}, w)$ such that, $P-$almost surely we have

$$\begin{cases} \nabla \cdot \left(\dfrac{\mathbf{k}(\mathbf{x}, w)}{\mu} \nabla p \right) = 0 & \\ q(\mathbf{x}, w) = -\dfrac{\mathbf{k}(\mathbf{x}, w) \nabla p}{\mu} & (\mathbf{x}, w) \ \in \Omega \times \mathbb{D} \\ p(\mathbf{x}, w) = p_0 & (\mathbf{x}, w) \ \in \partial\Omega_D^1 \times \mathbb{D} \\ \mathbf{q}(\mathbf{x}, w) \cdot \mathbf{n} = p_1 & (\mathbf{x}, w) \ \in \partial\Omega_N^1 \times \mathbb{D}, \end{cases} \tag{2.3}$$

where $\mathbf{k}(\mathbf{x}, w) = k(\mathbf{x}, w)\mathbf{I}_d$ is a random field with respect to a probability space $(\mathbb{D}, \mathbb{A}, P)$, $\mathbb{D}$ is the sample domain or event space, $\mathbb{A}$ is a $\sigma-$algebra on $\mathbb{D}$ and P is a probability measure. The permeability is generally taken to be log normal and therefore is given by

$$k(\mathbf{x}, \omega) = k_0(\mathbf{x}) + \delta e^{Y(\mathbf{x}, \omega)}, \tag{2.4}$$

where k_0 is a deterministic function, $\delta > 0$ is a constant and $Y(\mathbf{x}, \omega)$ is a Gaussian random field. If we think of k_0 as the reference permeability field, then $\delta e^{Y(\mathbf{x}, \omega)}$ is the random

variability to model the uncertainty in the description of k_0. We assume that k_0 satisfies the uniform lower bound

$$0 < k_1 \leq k_0(\mathbf{x}) \qquad \mathbf{x} \in \Omega. \tag{2.5}$$

This assumption is a reasonable because the permeability field is always positive. Since $\delta > 0$, it follows immediately that

$$0 < k_1 \leq k_0(\mathbf{x}) \leq k(\mathbf{x}, \omega) \qquad (\mathbf{x}, \omega) \in \Omega \times \mathbb{D}. \tag{2.6}$$

So the random permeability field $k(\mathbf{x}, \omega)$ is uniformly bounded from below. The Gaussian process is represented by the following Karhunen–Loeve expansion (2.7)

$$Y(\mathbf{x}, \omega) = \sum_{n=0}^{\infty} \sqrt{\lambda_n} f_n(\mathbf{x}) \mathcal{N}_n(\omega) \qquad (\mathbf{x}, \omega) \in \Omega \times \mathbb{D}, \tag{2.7}$$

where $\{\mathcal{N}_n\}$ is a family of independent Gaussian random variables with mean 0 and variance 1, and $\{\lambda_n, f_n\}$ are the eigenvalues and eigenfunctions of the covariance operator Q defined by

$$Qf(\mathbf{x}) = \int_{\Omega} C_r(\mathbf{x}, \mathbf{y}) f(\mathbf{y}) d\mathbf{y}, \qquad f \in L^2(\Omega),$$

where C_r is the covariance function. By definition of the covariance function, it is bounded, symmetric and positive definite. In geosciences the following exponential covariance function (kernel) is often used

$$C_r(\mathbf{x}, \mathbf{y}) = \sigma^2 \exp\left(-\sum_{i=1}^{d} \frac{|x_i - y_i|^2}{b_i^2} \right) \tag{2.8}$$

$$C_r(\mathbf{x}, \mathbf{y}) = \sigma^2 \exp\left(-\sum_{i=1}^{d} \frac{|x_i - y_i|}{b_i} \right), \tag{2.9}$$

where b_i is the the correlation length in i direction and σ^2 the variance. In either case, we have

$$\mathbf{E}[Y(\mathbf{x}, w)]^2 = \sigma^2 = C_r(\mathbf{x}, \mathbf{x}), \tag{2.10}$$

where $\mathbf{E}$ is the expectation.

In two dimensions the explicit formulae for the eigenvalues and eigenfunctions exist for the kernel (2.9) (see [27]). The well posedness of (2.2) and (2.3) in weak sense are

well known [28, 29] under the natural assumption that the tensor $\mathbf{k}$ is symmetric, positive definite and uniformly bounded, then there exist two positive constants k_1 and k_2 such that

$$k_1\zeta^T\zeta \leq \zeta^T\mathbf{k}^{-1}(\mathbf{x})\zeta \leq k_2\zeta^T\zeta \quad \forall \ \mathbf{x} \in \Omega \qquad \forall \zeta \in \mathbb{R}^d \tag{2.11}$$

for the deterministic system (2.2) and

$$k_1\zeta^T\zeta \leq \zeta^T\mathbf{k}^{-1}(\mathbf{x}, w)\zeta \leq k_2\zeta^T\zeta \qquad \text{a.e} \qquad (\mathbf{x}, w) \in \Omega \times \mathbb{D} \qquad \forall \zeta \in \mathbb{R}^d \tag{2.12}$$

for the stochastic system (2.3), which excludes the use of the unbounded random variables such as Gaussian random variable in the Karhunen–Loeve expansion (2.7). If $\mathbf{k}$ is symmetric, positive definite and uniformly bounded from below, the well posedness is ensured, in this case the solution belongs in a functional space depending on $\mathbf{k}$ (see [30]). For well posedness we also need that $p_0 \in H^{1/2}(\partial\Omega_D^1)$, where the Sobolev space $H^{1/2}(\partial\Omega_D^1)$ is viewed as

$$H^{1/2}(\partial\Omega_D^1) = \left\{ f : \ f = w|_{\partial\Omega_D^1} \text{for some } w \in H^1(\Omega) \cap C(\overline{\Omega}) \right\}.$$

The space $C(\overline{\Omega})$ is the set of continuous functions defined in $\overline{\Omega}$.

2.2 Mild solution for advection-diffusion-reaction

As we are sure of the existence and uniqueness of the Darcy velocity field $\mathbf{q}$, let us consider the nonlinear (ADR) model which can be formulated as following: Find the concentration $X = X(\mathbf{x}, t)$ such that

$$\begin{cases} \partial X/\partial t + AX = R(\mathbf{x}, t, X) & (\mathbf{x}, t) \in \Omega \times [0, T] \\ X(\mathbf{x}, 0) = X_0 & \mathbf{x} \in \Omega \\ X(\mathbf{x}, t) = X_1(\mathbf{x}, t) & (\mathbf{x}, t) \in \partial\Omega_D \times [0, T] \\ \gamma_A X(\mathbf{x}, t) = X_2(\mathbf{x}, t) & (\mathbf{x}, t) \in \partial\Omega_N \times [0, T] \end{cases} \tag{2.13}$$

where

$$\begin{aligned} AX \equiv A(\mathbf{x}, X) &= -\nabla \cdot (\mathbf{D}\nabla X) + \nabla \cdot (\mathbf{q}(\mathbf{x})X) \\ &= -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(D_{i,j}(\mathbf{x})\frac{\partial X}{\partial x_j}\right) + \sum_{i=1}^{d} q_i(\mathbf{x})\frac{\partial X}{\partial x_i} + (\nabla \cdot \mathbf{q})\, X \\ &= -\sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(D_{i,j}(\mathbf{x})\frac{\partial X}{\partial x_j}\right) + \sum_{i=1}^{d} q_i(\mathbf{x})\frac{\partial X}{\partial x_i} \end{aligned}$$

and

$$\gamma_A X \equiv \frac{\partial X}{\partial \nu_A} = \sum_{i=1}^{d} n_i(\mathbf{x}) D_{i,j}(\mathbf{x}) \frac{\partial X}{\partial x_i}$$

$\mathbf{n} = (n_i)$ is the unit outward normal to $\partial\Omega_N$ with $\partial\Omega = \partial\Omega_D \cup \partial\Omega_N$.

Here we focus on the homogeneous case ($X_1 = X_2 = 0$) as the general case can be put in the homogeneous form using a trace operator [11]. Having the homogeneous boundary conditions for our model problem (2.8) implies that the function R contains extra terms from boundary conditions of the initial problem. We have also scaled the original equation from the advection–diffusion–reaction model by the porosity function ϕ. For well posedness of (2.8), we make the following assumptions:

Assumption 2.1 ***[Ellipticity condition of the diffusion tensor]***

We assume that $\mathbf{D}$ *is symmetric,* $D_{i,j} \in L^\infty(\Omega)$ *and there exists a positive constant* $c_1 > 0$ *such that*

$$\sum_{i,j=1}^{d} D_{i,j}(\mathbf{x})\xi_i\xi_j \geq c_1|\xi|^2 \qquad \forall \xi \in \mathbb{R}^d \quad \mathbf{x} \in \overline{\Omega} \quad c_1 > 0. \tag{2.14}$$

Assumption 2.2 ***[Lipschitz condition for nonlinear reaction term]***

The nonlinear R *is continuous and satisfies a local Lipschitz condition with respect to the variable* X*, i.e. there exists a positive constant* $L > 0$ *such that*

$$|R(\mathbf{x}, t, u) - R(\mathbf{x}, t, v)| \leq L\left(1 + |u|^\gamma + |v|^\gamma\right)|u - v| \quad \forall u, v \in \mathbb{R} \;\; x \in \overline{\Omega},\; t \in [0, T], \tag{2.15}$$

with $\gamma = 2$ *for* $d = 3$ *and* $\gamma \in [0, \infty)$ *for* $d = 2$.

We introduce the space V that depends on the choice of the boundary conditions. For full Dirichlet boundary conditions we let

$$V = H_0^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \text{ on } \partial\Omega\},$$

and for Neumann and mixed boundary conditions $V = H^1(\Omega)$. The bilinear form associated with the operator A is given by

$$a(u,v) = \int_\Omega \left(\sum_{i,j=1}^{d} D_{i,j} \frac{\partial u}{\partial x_j} \frac{\partial v}{\partial x_i} + \sum_{i=1}^{d} q_i \frac{\partial u}{\partial x_j} v \right) dx \qquad u, v \in V. \tag{2.16}$$

Coercivity of the bilinear form (2.16) is answered in part by the following theorem.

Theorem 2.3 ***[Gårding's inequality] [31]***

Assume that Assumption 2.1 holds and $q_i \in L^\infty(\Omega)$, then there exists a positive constant c_0 such that

$$a(v,v) + c_0\|v\|^2 \geq \frac{c_1}{2}\|v\|^2_{H^1(\Omega)} \qquad \forall v \in V. \tag{2.17}$$

Proof. The proof is mostly given for $q_i \in C^1(\Omega)$ with the choice

$$c_0 = \sup_{x\in\Omega} \frac{1}{2} \left(\sum_{i=1}^{d} \frac{\partial q_i}{\partial x_i} \right)$$

(see [31]).

Here we give a proof for $q_i \in L^\infty(\Omega)$. Let $c > 0$. For $v \in V$ by the ellipticity condition (2.14) we have

$$a(v,v) + c\|v\|^2_{L^2(\Omega)} \;\geq\; c_1 \, |v|_{H^1(\Omega)} + \int_\Omega \left(\sum_{i=1}^{d} q_i \frac{\partial v}{\partial x_i} v + cv^2 \right) d\mathbf{x}, \tag{2.18}$$

where $|.|_{H^1(\Omega)}$ is the semi norm of the Sobolev space $H^1(\Omega)$, the so called $H_0^1(\Omega)-$norm or gradient norm in the space $H_0^1(\Omega)$. By Hölder's inequality we have

$$\left| \int_\Omega \sum_{i=1}^{d} q_i \, \frac{\partial v}{\partial x_i} \, v \, d\mathbf{x} \right| \;\leq\; \int_\Omega \sum_{i=1}^{d} |q_i| \, \left|\frac{\partial v}{\partial x_i}\right| \, |v| \, d\mathbf{x} \tag{2.19}$$

$$\leq \sum_{i=1}^{d} \|q_i\|_{L^\infty(\Omega)} \left\| \frac{\partial v}{\partial x_i} \right\| \, \|v\| \tag{2.20}$$

$$\leq \left(\sum_{i=1}^{d} \|q_i\|^2_{L^\infty(\Omega)} \right)^{1/2} |v|_{H^1(\Omega)} \|v\|. \tag{2.21}$$

Set

$$\beta = \left(\sum_{i=1}^{d} \|q_i\|^2_{L^\infty(\Omega)} \right)^{1/2}, \tag{2.22}$$

then

$$a(v,v) + c\|v\|^2 \;\; \geq \;\; c_1 \, |v|^2_{H^1(\Omega)} - \beta |v|_{H^1(\Omega)} \|v\| + c\|v\|^2. \tag{2.23}$$

Recall Young's inequality which states that if a and b are nonnegative real numbers and p and q are positive real numbers such that $1/p + 1/q = 1$ then

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}. \tag{2.24}$$

Using this inequality with $a = \sqrt{c_1}|v|_{H^1(\Omega)}, \; b = \frac{\beta}{\sqrt{c_1}}\|v\|, \; p = q = 1/2$ yields

$$a(v,v) + c\|v\|^2 \;\; \geq \;\; \frac{c_1}{2}|v|^2_{H^1(\Omega)} + \left(c - \frac{\beta^2}{2c_1} \right) \|v\|^2, \tag{2.25}$$

Taking $c = c_0$ such that

$$c_0 - \frac{\beta^2}{2c_1} \geq \frac{c_1}{2} \Leftrightarrow c_0 \geq \frac{\beta^2}{2c_1} + \frac{c_1}{2}, \tag{2.26}$$

yields

$$a(v,v) + c_0\|v\|^2 \geq \; \frac{c_1}{2} \left(|v|^2_{H^1(\Omega)} + \|v\|^2 \right) = \frac{c_1}{2}\|v\|^2_{H^1(\Omega)}. \tag{2.27}$$

■

By adding $c_0 X$ to both sides of the first equation of (2.8) we have a new operator that we still call A corresponding to the new bilinear form that we still call a such that the following coercivity property holds

$$a(v,v) \geq \; \frac{c_1}{2}\|v\|^2_{H^1(\Omega)} \qquad \forall v \in V. \tag{2.28}$$

We will still call the right hand side of the first equation of (2.8) R, it is obvious that the new R also satisfies the local Lipschitz condition (2.15). Furthermore with a slight abuse of notation R will denote the nonlinear operator $X \to R(\cdot, \cdot, X)$.

Theorem 2.4 *Suppose that the reaction term R satisfies Assumption 2.2. For each bounded set $\mathcal{B} \subset V$ there is a constant $C(\mathcal{B})$ such that*

$$\|R(u) - R(v)\|_{H^{-1}(\Omega)} \leq C(\mathcal{B})\|u - v\|, \qquad \forall u, v \in \mathcal{B} \tag{2.29}$$

$$\|R(u) - R(v)\| \leq C(\mathcal{B})\|u - v\|_{H^1(\Omega)}, \qquad \forall u, v \in \mathcal{B} \; . \tag{2.30}$$

Proof. See [32]. ■

By Green's formula we have

$$a(u,v) = (Au,v) \qquad \forall u \in \mathbb{H} \cap H^2(\Omega) = \mathcal{D}(A) \qquad \forall v \in V, \tag{2.31}$$

where

$$\mathbb{H} = \{v \in H^1(\Omega) : v = 0 \ \text{ on } \ \partial\Omega_D\} \tag{2.32}$$

and $\mathcal{D}(A)$ the domain of the operator A. Therefore the weak form of (2.8) is to find the function $X(t) \in \mathcal{D}(A)$ such that

$$\begin{cases} (X_t, \chi) + (AX, \chi) = (R(X), \chi) & \forall \chi \in V, \quad t \in [0,T] \\ X(t) = X_0. \end{cases} \tag{2.33}$$

The $V-$ellipticity (2.28) implies that $-A$ is a sectorial on $L^2(\Omega)$ (see [12, 17]) i.e. there exists C_1, $\theta \in (\frac{1}{2}\pi, \pi)$ such that

$$\|(\lambda I + A)^{-1}\|_{L(L^2(\Omega))} \leq \frac{C_1}{|\lambda|} \qquad \lambda \in S_\theta, \tag{2.34}$$

where $S_\theta = \big\{\lambda \in \mathbb{C} : \lambda = \rho e^{i\phi},\ \rho > 0,\ 0 \leq |\phi| \leq \theta\big\}$.

Then $-A$ is the infinitesimal generator of bounded analytic semigroups $S(t) := e^{-tA}$ on $L^2(\Omega)$ such that

$$S(t) := e^{-tA} = \frac{1}{2\pi i} \int_{\mathcal{C}} e^{t\,\lambda} (\lambda I + A)^{-1} d\lambda, \qquad t > 0 \tag{2.35}$$

where $\mathcal{C}$ denotes a path that surrounds the spectrum of $-A$. By Duhamel's principle we may represent solutions of (2.8) by the following integral equation

$$X(t) = S(t)X_0 + \int_0^t S(t-s)R(s, X(s))ds, \qquad t \in [0,T]\,. \tag{2.36}$$

We may now apply a standard argument to obtain local existence and uniqueness.

Theorem 2.5 *Under Assumptions 2.1 and 2.2, assume that $q_i \in L^\infty(\Omega)$. For any bounded set $\mathcal{B}_0 \subset V$ there is $t^* = t^*(\mathcal{B}_0)$ such that equation (2.36) has a unique solution $X \in C([0,t^*], H^1(\Omega))$ for any $X_0 \in \mathcal{B}_0$.*

Proof. The proof is based on applying the contraction mapping principle in the topology of the Banach space $C([0,t^*],H^1(\Omega))$ to the integral equation (2.36) [12, Theorem 3.3.3] or [33, Theorem 6.3.1]. ■

The coercivity property implies that the set of the real part of the spectrum of A is non negative, which allows the definition of the fractional power of A as: for any $\alpha > 0$

$$\begin{cases} A^{-\alpha} = \dfrac{1}{\Gamma(\alpha)} \int_0^\infty t^{\alpha-1} e^{-At} dt \\ \\ A^{\alpha} = (A^{-\alpha})^{-1} \end{cases} \tag{2.37}$$

where $\Gamma(\alpha)$ is the Gamma function of α [12]. We denote by $\|.\|_\alpha := \|A^{\alpha/2}.\|$ the norm of the space $\mathcal{D}(A^{\alpha/2})$.

We recall some basic properties of the semigroup $S(t)$ generated by $-A$.

Proposition 2.6 ***[Smoothing properties of the semi group [12]]***
Let $\beta \geq 0$ and $0 \leq \gamma \leq 1$, then there exists $C > 0$ such that

$$\|A^\beta S(t)\|_{L(L^2(\Omega))} \leq Ct^{-\beta} \quad for \quad t > 0,$$

$$\|A^{-\gamma}(I - S(t))\|_{L(L^2(\Omega))} \leq Ct^{\gamma} \quad for \quad t \geq 0.$$

In addition, the following results hold

$$A^\beta S(t) = S(t)A^\beta \quad on \quad \mathcal{D}(A^\beta).$$

$$If \quad \beta \geq \gamma \quad then \quad \mathcal{D}(A^\beta) \subset \mathcal{D}(A^\gamma).$$

$$\|D_t^l S(t)v\|_\beta \leq Ct^{-l-(\beta-\alpha)/2} \|v\|_\alpha, \ \ t > 0, \ v \in \mathcal{D}(A^{\alpha/2}) \ \ l = 0, 1,$$

where $D_t^l := \dfrac{d^l}{dt^l}$.

2.3 A cell–centred finite volume for ADR

A cell–centred finite volume methods for heterogeneous and anisotropic diffusion problems remains a challenging problem. An active area of research is to make the approximation of the diffusion flux more efficient and simple as possible (see [34] for the references). The finite volume method is widely applied when the differential equations are in divergence form. To

obtain a finite volume discretization, the domain Ω is subdivided into N subdomains $(A_i)_{i\in\mathcal{I}}$ called control volumes or control domains such that the collection of all those subdomains form a partition of Ω. The common feature of all finite volume methods is to integrate the equation over each control volume $A_i,\ i \in \mathcal{I}$ and apply Gauss's divergence theorem to convert the volume integral to a surface integral. For our parabolic problem (2.8), finite volume methods differ in the way they approximate the diffusion flux in the case of a full diffusion tensor (dispersion tensor). Here we present one way to approximate the diffusion tensor given in [35].

2.3.1 A cell–centred finite volume space discretization in an admissible mesh for full diffusion tensor

Here we describe a finite volume method as in Eymard and al. [35] for heterogeneous and anisotropic diffusion problems. Denote by $\{\mathbf{x}_i\}_{i\in\mathcal{I}}$ a consecutively numbered set of points in Ω that includes all vertices of Ω. $\mathcal{I}$ is the corresponding set of indices. Standard Voronoi meshes $\mathcal{T} = \{A_i\}_{i\in\mathcal{I}}$ are defined as follows

$$A_i = \{\mathbf{x} \in \Omega \ \mid \mathbf{x} - \mathbf{x}_i \mid < \mid \mathbf{x} - \mathbf{x}_j \mid \text{ for all } i \neq j\} \cap \Omega, \quad i \in \mathcal{I}, \tag{2.38}$$

where $\mid \mathbf{x} - \mathbf{y} \mid$ denotes the Euclidean distance between $\mathbf{x}$ and $\mathbf{y}$. If near every point $\mathbf{x}_i,\ i \in \mathcal{I}$ a local inner product with the corresponding norm $\mid . \mid_i$ is given, we can generalize (2.38) by defining the control volumes $\mathcal{T} = \{A_i\}_{i\in\mathcal{I}}$ as

$$A_i = \{\mathbf{x} \in \Omega \ \mid \mathbf{x} - \mathbf{x}_i \mid_i < \mid \mathbf{x} - \mathbf{x}_j \mid_j \text{ for all } i \neq j\} \cap \Omega, \quad i \in \mathcal{I}. \tag{2.39}$$

The admissible meshes $\mathcal{T}$ in [35, Definition 3.8] for problem (2.8) with the full diffusion tensor $\mathbf{D}$ is defined in the opposite way to the Voronoi meshes. We summarize [35, Definition 3.8] here.

Definition 2.7 ***[An admissible mesh]***

An admissible mesh $\mathcal{T}$ for problem (2.8) with the full diffusion tensor $\mathbf{D}$ is given by:

- *A set $\{A_i\}_{i\in\mathcal{I}}$ of control volumes such that $\overline{\Omega} = \bigcup_{i\in\mathcal{I}} \overline{A_i}$ with the corresponding local inner product induced by $\mathbf{D}_{A_i}^{-1}$ where*

$$\mathbf{D}_{A_i} = \frac{1}{\text{mes}(A_i)} \int_{A_i} \mathbf{D}(\mathbf{x}) d\mathbf{x}.$$

- *The corresponding set of center points* $\{\mathbf{x}_i\}_{i\in\mathcal{I}}$ *such that*

  (a) $\mathbf{x}_i \in \overline{A}_i,\ i \in \mathcal{I}$.

  (b) $\mathbf{x}_i$ *is the intersection of the straight lines perpendicular to the boundary of* A_i *with respect to the inner product induced by* $\mathbf{D}_{A_i}^{-1}$.

Let h be the maximum mesh size of $\mathcal{T}$. We denote by $\mathcal{T}_h$ a dual Delaunay triangulation of $\mathcal{T}$. $\{\mathbf{x}_i\}_{i\in\mathcal{I}}$ is the set of vertices of $\mathcal{T}_h$.

Let us illustre Definition 2.7 to make it more understandable.

Example 2.8

- *In the case where the diffusion tensor* $\mathbf{D}$ *is diagonal and* Ω *is a rectangular or parallelepiped domain, any rectangular grid* ($d = 2$) *or parallelepiped grid* ($d = 3$) *is an admissible mesh. The set* $\{\mathbf{x}_i\}$ *is the set of centers of gravity of the rectangular grid or parallelepiped grid. The inner product induced locally by* $\mathbf{D}_{A_i}^{-1}$ *is equivalent to the standard inner product corresponding to the Euclidean norm* $|.|$. *This mesh yields a 5-point scheme* ($d = 2$) *and 7-point scheme* ($d = 3$) *for our model problem .*

- *If* $d = 2$, *for isotropic and heterogeneous media* ($\mathbf{D}(\mathbf{x}) = b(\mathbf{x})I_2$ *a. e* $\mathbf{x} \in \Omega$) *we can define a triangular admissible mesh* $\mathcal{T}$ *to be a family of open triangular disjoint subsets of* Ω *such that two triangles having a common edge have also two common vertices. The angles of the triangles are assumed to be less than* $\frac{\pi}{2}$ *to allow the orthogonal bisectors to intersect inside each triangle, thus naturally defining the center point* x_i *of the control volume* A_i. *The finite volume scheme defined on such mesh yields a 4-point scheme for our model problem . The inner product induced locally by* $\mathbf{D}_{A_i}^{-1}$ *is equivalence to the standard inner product corresponding to the Euclidean norm* $|.|$.

To make notation easier, throughout this thesis, we will identify $\mathcal{T}$ to $\mathcal{I}$, then to say $A_i \in \mathcal{T}$ we will say $i \in \mathcal{T}$.

Consider the modified model problem of (2.8) where $c_0 X$ is added on both sides of the first equation of problem (2.8) without scaling by the porosity ϕ, where c_0 is defined in Theorem 2.17. Consider an admissible mesh $\mathcal{T}$ in the sense of Definition 2.7. Denote by $\mathcal{E}$ the set of edges of control volume of $\mathcal{T}$, $\mathcal{E}_{int}$ the set of interior edges of control volume of $\mathcal{T}$, $X_i(t)$ the approximation of X at time t at the center (or at any point) of the control

volume $i \in \mathcal{T}$ and $X_\sigma(t)$ the approximation of X at time t at the center (or at any point)of the edge $\sigma \in \mathcal{E}$. For a control volume $i \in \mathcal{T}$, denote by $\mathcal{E}_i$ the set of edges of i, mes(i) the Lebesgue measure of the control volume $i \in \mathcal{T}$.

As in [11, 35], integration over any control volume $i \in \mathcal{T}$, using the divergence theorem to convert the integral over i to a surface integral, finite differences for the diffusion flux approximation [35] and the upwind technique for the advection flux approximation yields

$$\begin{cases} \operatorname{mes}(i)\phi_i \dfrac{dX_i(t)}{dt} + \displaystyle\sum_{\sigma\in\mathcal{E}_i} (F_{i,\sigma}(t) + q_{i,\sigma}X_{\sigma,+}(t)) + c_0 \operatorname{mes}(i) X_i(t) = \operatorname{mes}(i)\, R(\mathbf{x}_i, t, X_i(t)), \\ \phi_i = \dfrac{1}{\operatorname{mes}(i)} \int_i \phi(\mathbf{x}) d\mathbf{x}, \qquad D_{i,\sigma} = |\mathbf{D}_i\, \mathbf{n}_{i,\sigma}|, \quad \mathbf{D}_i = \dfrac{1}{\operatorname{mes}(i)} \int_i \mathbf{D}(\mathbf{x}) d\mathbf{x}, \\ F_{i,\sigma}(t) = \operatorname{mes}(\sigma)\; D_{i,\sigma} \dfrac{X_\sigma(t) - X_i(t)}{d_{i,\sigma}}, \qquad\qquad \sigma \subsetneq \partial\Omega_N \\ F_{i,\sigma}(t) = \dfrac{1}{\operatorname{mes}(\sigma)} \int_\sigma X_2(\mathbf{x}, t) d\sigma, \qquad\qquad \sigma \subseteq \partial\Omega_N \\ q_{i,\sigma} = \int_\sigma \mathbf{q} \cdot \mathbf{n}_{i,\sigma} d\sigma \qquad\qquad \forall i \in \mathcal{T}, \;\; \forall \sigma \in \mathcal{E}_i. \end{cases} \tag{2.40}$$

Here $\mathbf{n}_{i,\sigma}$ is the normal unit vector to σ outward to i, mes(σ) is the Lebesgue measure of the edge $\sigma \in \mathcal{E}_i$.

Since the flux is continuous at the interface of two control volumes i and j (denoted by $i \mid j$) we therefore have $F_{i,\sigma}(t) = -F_{j,\sigma}(t)$ for $\sigma = i \mid j$, which yields

$$\begin{cases} \tau_\sigma = \operatorname{mes}(\sigma) \dfrac{D_{i,\sigma} D_{j,\sigma}}{D_{i,\sigma}\, d_{i,\sigma} + D_{j,\sigma} d_{j,\sigma}} \quad (\text{transmissibility through } \sigma) \\ F_{i,\sigma}(t) = -\tau_\sigma \left(X_j(t) - X_i(t)\right) = -\dfrac{\mu_\sigma \operatorname{mes}(\sigma)}{d_{i,j}} \left(X_j(t) - X_i(t)\right), \qquad \sigma = i \mid j \end{cases} \tag{2.41}$$

with

$$\mu_\sigma = d_{i,j} \frac{D_{i,\sigma} D_{j,\sigma}}{D_{i,\sigma}\, d_{i,\sigma} + D_{j,\sigma} d_{j,\sigma}}. \tag{2.42}$$

where $d_{i,j}$ is the distance between the center of i and center of j and $d_{i,\sigma}$ the distance between the center of i and the edge σ. We will also denote by d_σ the distance $d_{i,j}$ or $d_{i,\sigma}$ for $\sigma = i \mid j$ or $\sigma = \mathcal{E}_i \cap \partial\Omega$ respectively.

For $\sigma \subset \partial\Omega_D$ assuming that $X_1 \in C(\partial\Omega_D)$, we can also write

$$\begin{aligned} F_{i,\sigma}(t) &= \text{mes}(\sigma)\, D_{i,\sigma} \frac{X_\sigma(t) - X_i(t)}{d_{i,\sigma}} \\ &= -\tau_\sigma \left(X_j(t) - X_i(t)\right) \\ &= -\frac{\text{mes}(\sigma)\mu_\sigma}{d_{i,\sigma}} \left(X_j(t) - X_i(t)\right) \end{aligned}$$

with

$$\begin{cases} X_j(t) = X_\sigma(t) = X_1(\mathbf{x}_\sigma, t) \\ \\ \tau_\sigma = \dfrac{\text{mes}(\sigma) D_{i,\sigma}}{d_{i,\sigma}} \\ \\ \mu_\sigma = D_{i,\sigma} \end{cases} \tag{2.43}$$

and $\mathbf{x}_\sigma$ the center of σ.

The upwind term for advection flux $X_{\sigma,+}$ is defined as

$$X_{\sigma,+}(t) = \begin{cases} X_i(t) \quad \text{if} \quad q_{i,\sigma} \geqslant 0 \\ \\ X_j(t) \quad \text{if} \quad q_{i,\sigma} < 0 \end{cases} \quad \text{for} \quad \sigma = i \mid j \tag{2.44}$$

$$X_{\sigma,+}(t) = \begin{cases} X_i(t) \quad \text{if} \quad q_{i,\sigma} \geqslant 0 \\ \\ X_\sigma(t) \quad \text{if} \quad q_{i,\sigma} < 0 \end{cases} \quad \text{for} \;\; \sigma \in \mathcal{E}_i \cap \partial\Omega. \tag{2.45}$$

We can write $X_{\sigma,+}$ as

$$X_{\sigma,+} = r_\sigma X_i(t) + (1 - r_\sigma) X_j(t), \quad \sigma = i \mid j \tag{2.46}$$

where $r_\sigma = \dfrac{1}{2}(\text{sign}(q_{i,\sigma}) + 1)$. The finite volume space discretization for the model problem

(2.8) is given by

$$\begin{cases} \operatorname{mes}(i)\dfrac{dX_i(t)}{dt} + \displaystyle\sum_{\sigma=i|j\in\mathcal{E}_i}\left(-\frac{\operatorname{mes}(\sigma)\,\mu_\sigma}{d_{i,j}}\left(X_j(t)-X_i(t)\right)+q_{i,\sigma}\left(r_\sigma X_i(t)+(1-r_\sigma)X_j(t)\right)\right) \\ \\ + \displaystyle\sum_{\sigma\in\mathcal{E}_i\cap\partial\Omega_D}\left(\frac{\operatorname{mes}(\sigma)\,\mu_\sigma}{d_{\sigma,i}}X_i(t)+q_{i,\sigma}r_\sigma X_i(t)\right)+\sum_{\sigma\in\mathcal{E}_i\cap\partial\Omega_N}q_{i,\sigma}X_i(t)+c_0\operatorname{mes}(i)X_i \\ \\ = \operatorname{mes}(i)\,R(X_i(t)), \qquad\qquad \forall i\ \in\mathcal{T} \end{cases} \tag{2.47}$$

where here $R(X_i(t)) := R(\mathbf{x}_i, t, X_i(t))$ is the initial reaction plus boundary conditions contribution. Notice that in (2.47), the only unknowns are the center values. We have eliminated the values at the Neumann boundary in the advection flux by the relation

$$\begin{aligned} F_{i,\sigma}(t) &= \frac{1}{\operatorname{mes}(\sigma)}\int_\sigma X_2(\mathbf{x},t)d\gamma(\mathbf{x}) \approx -\tau_\sigma\left(X(\mathbf{x}_\sigma,t)-X_i(t)\right), \qquad \sigma\subset\partial\Omega_N \\ \tau_\sigma &= \frac{\operatorname{mes}(\sigma)D_{i,\sigma}}{d_{i,\sigma}}, \end{aligned}$$

which implies

$$X_{\sigma,+} = r_\sigma X_i(t) + (1-r_\sigma)X(\mathbf{x}_\sigma,t) = X_i(t) + (1-r_\sigma)\tau_\sigma^{-1}F_{i,\sigma}(t), \qquad \sigma\subset\partial\Omega_N.$$

The scheme (2.47) clearly indicates the affinity of the finite volume method to the finite difference method. However, for the subsequent analysis it is more convenient to rewrite scheme (2.47) in a discrete variational form. Multiplying equation (2.47) by arbitrary numbers $v_i \in \mathbb{R}$ and summing the results over all control volume in $\mathcal{T}$ yields

$$\begin{cases} \displaystyle\sum_{i\in\mathcal{T}}\left[\operatorname{mes}(i)\frac{dX_i(t)}{dt} + \sum_{\sigma=i|j\in\mathcal{E}_i}\frac{\operatorname{mes}(\sigma)\,\mu_\sigma}{d_{i,j}}\left(X_i(t)-X_j(t)\right)+q_{i,\sigma}\left(r_\sigma X_i(t)+(1-r_\sigma)X_j(t)\right)\right. \\ \\ \left. + \displaystyle\sum_{\sigma\in\mathcal{E}_i\cap\partial\Omega_D}\left(\frac{\operatorname{mes}(\sigma)\,\mu_\sigma}{d_{\sigma,i}}X_i(t)+q_{i,\sigma}r_\sigma X_i(t)\right)+\sum_{\sigma\in\mathcal{E}_i\cap\partial\Omega_N}q_{i,\sigma}X_i(t)+c_0\operatorname{mes}(i)X_i\right]v_i \\ \\ = \displaystyle\sum_{i\in\mathcal{T}}\operatorname{mes}(i)\,R(X_i(t))v_i. \end{cases} \tag{2.48}$$

Let $V_h \subset V$ denote the space of continuous functions that are piecewise linear over the Delaunay triangulation $\mathcal{T}_h$ (dual of $\mathcal{T}$), then the values $X_i(t)$ and v_i can be interpolated in

V_h. There are unique functions $X_h(t), v_h \in V_h$ such that $X_h(t)(\mathbf{x}_i) = X_i(t)$ and $v_h(\mathbf{x}_i) = v_i$ for all $i \in \mathcal{T}$, where $\mathbf{x}_i$ is a center of the control volume $i \in \mathcal{T}$ ($\mathbf{x}_i$ is also a vertex in $\mathcal{T}_h$).

Denote by a_h the bilinear form defined by

$$\begin{cases} a_h(u_h, v_h) = \sum_{i\in\mathcal{T}} \sum_{\sigma=i|j\in\mathcal{E}_i} \left(-\dfrac{\mathrm{mes}(\sigma)\,\mu_\sigma}{d_{i,j}} (u_j - u_i) + q_{i,\sigma} (r_\sigma u_i + (1 - r_\sigma) u_j) \right) v_i \\ \\ + \sum_{i\in\mathcal{T}} \left(\sum_{\sigma\in\mathcal{E}_i\cap\partial\Omega_D} \left(\dfrac{\mathrm{mes}(\sigma)\,\mu_\sigma}{d_{\sigma,i}} u_i + q_{i,\sigma} r_\sigma u_i \right) + \sum_{\sigma\in\mathcal{E}_i\cap\partial\Omega_N} q_{i,\sigma} u_i + c_0\,\mathrm{mes}(i) u_i \right) v_i, \\ \\ \forall u_h, v_h \in V_h, \end{cases} \tag{2.49}$$

and by $\langle ., . \rangle_{0,h}$ the scalar product on $C(\overline{\Omega}) \supset V_h$ defined by

$$\langle u, v \rangle_{0,h} = \sum_{i\in\mathcal{T}} \mathrm{mes}(i) u_i v_i, \quad u_i = u(\mathbf{x}_i), \quad v_i = v(\mathbf{x}_i), \qquad u, v \in C(\overline{\Omega}), \tag{2.50}$$

the corresponding norm is the discrete $L^2(\Omega)$ norm denoted by $\|.\|_{0,h}$. It is proved in [11] that $\|.\|_{0,h}$ is equivalent to the $L^2(\Omega)$ norm $\|.\|$ when the mesh $\mathcal{T}$ is regular in V_h.

Previous results allow us to write the following variational form of our finite volume scheme (2.48).

$$\begin{cases} \langle \dfrac{d}{dt} X_h, \varphi \rangle_{0,h} + a_h(X_h(t), \varphi) = \langle R(X_h(t)), \varphi \rangle_{0,h}, \quad \forall \varphi \in V_h, \quad t \in (0, T], \\ \\ X_h(0) = X_{h\,0}, \end{cases} \tag{2.51}$$

with

$$R(X_h(t)(\mathbf{x}_i) = R(X_i(t)) := R(\mathbf{x}_i, t, X_i(t)) = R(\mathbf{x}_i, t, X_h(t)(\mathbf{x}_i)), \qquad \forall i \in \mathcal{T}.$$

Consider the operator $A_h : V_h \to V_h$ such that

$$\langle A_h \psi, \chi \rangle_{0,h} = a_h(\psi, \chi) \quad \forall \psi, \chi \in V_h. \tag{2.52}$$

The semidiscrete solution in V_h is then given by: Find $X_h(t) \in V_h$ such that

$$\begin{cases} \dfrac{dX_h}{dt} + A_h X_h = P_h R(X_h) \qquad t \in (0, T] \\ \\ X_h(0) = X_{0h} \end{cases} \tag{2.53}$$

where P_h is the orthogonal projection defined from $u \in C(\overline{\Omega})$ to V_h by

$$\langle P_h u, \chi \rangle_{0,h} = \langle u, \chi \rangle_{0,h} \quad \forall \chi \in V_h. \tag{2.54}$$

Remark 2.9 *To obtain the integral of the Darcy velocity ($q_{i,\sigma}$) using in (2.49), the admissible $\mathcal{T}$ for ADR in Definition 2.7 need to be also admissible for the pressure–velocity system (2.2), and the corresponding finite volume scheme is given by*

$$\begin{cases} \sum\limits_{\sigma \in \mathcal{E}_i} q_{i,\sigma} = 0, \\ k_{i,\sigma} = |\mathbf{k}_i \mathbf{n}_{i,\sigma}|, \quad \mathbf{k}_i = \dfrac{1}{\text{mes}(i)} \int_i \mathbf{k}(\mathbf{x}) d\mathbf{x}, \\ q_{i,\sigma} = \dfrac{\text{mes}(\sigma) k_{i,\sigma} k_{j,\sigma}}{\mu \ (k_{i,\sigma}\ d_{i,\sigma} + k_{j,\sigma} d_{j,\sigma})} \left(p_j - p_i\right), & \sigma = i|j \in \mathcal{E}_{int} \\ q_{i,\sigma} = \dfrac{\text{mes}(\sigma) k_{i,\sigma}}{\mu\, d_{i,\sigma}} \left(p_0(\mathbf{x}_\sigma) - p_i\right), & \sigma \subset \partial\Omega_D^1 \\ q_{i,\sigma} = \int_\sigma p_1(\mathbf{x}, t) d\sigma & \sigma \subseteq \partial\Omega_N^1. \end{cases} \tag{2.55}$$

2.4 Standard time discretizations for ADR

We briefly describe two standard time-stepping schemes, the implicit Euler scheme and the semi implicit Euler scheme. Later we use these for comparison with the exponential scheme of order one, (ETD1). Given the initial data $X_h^0 = X^0$, the implicit Euler scheme for the system (2.53) is

$$\frac{X_h^{n+1} - X_h^n}{\Delta t} = -A_h X_h^{n+1} + P_h R(X_h^{n+1}, t_{n+1}), \qquad t_n = n\Delta t \tag{2.56}$$

and the semi implicit scheme is

$$\frac{X_h^{n+1} - X_h^n}{\Delta t} = -A_h X_h^{n+1} + P_h R(X_h^n, t_n) \tag{2.57}$$

where $\Delta t = t_{n+1} - t_n$ is the fixed time-step. For the implicit Euler method we have to solve a non-linear algebraic equation of the form

$$f(Y) = (I + \Delta t\ A_h)Y - \Delta t\, P_h R(Y, t_{n+1}) - X_h^n = 0$$

at each time-step. For brevity we denote X_h^{n+1} as Y. We use Newton's method and a variant of Newton's method designed for semi-linear problems [11]. We solve the linear systems using the standard backslash solver in Matlab$^{\text{TM}}$at each iteration in the exact Newton's method. For the variant of Newton's method, the Jacobian of f, $J(Y)$, is approximated by its constant linear part so that $J(Y) \approx I + \Delta t\, A_h$. The corresponding quasi-Newton iteration is then given by

$$\begin{aligned} Y_{k+1} &= Y_k - (I + \Delta t\, A_h)^{-1} f(Y_k) \\ &= (I + \Delta t\, A_h)^{-1} (\Delta t\, P_h R(Y_k, t_{n+1}) + X_h^n). \end{aligned}$$

This is equivalent to a fixed point method to solve the equivalent equation

$$(I + \Delta t\, A_h)^{-1} f(Y) = 0.$$

The approximation of the Jacobian by its constant linear part allows us to compute the matrix factorisation only once and to reuse this at each time-step. In the quasi-exact Newton's method and the semi implicit Euler scheme we solve the linear systems using either an LU-decomposition or the standard solver in Matlab$^{\text{TM}}$.

2.5 Iterative linear solvers

2.5.1 Affine linear iterative methods

Implementation of the standard time integrators means solving at each time step a linear system of the form

$$\mathbf{BX} = \mathbf{b}, \tag{2.58}$$

where $\mathbf{B} \in \mathbb{R}^{n\times n}$, $\mathbf{X}, \mathbf{b} \in \mathbb{R}^n, n \in \mathbb{N}$. Many soflware packages have some direct solvers already implemented for the linear system (2.58), which are generally not practical for large size problems. Among iterative solvers the simple class is the so called affine-linear iterative methods, where the solution of (2.58) is the fixed-point of the function

$$\Phi(\mathbf{X}) = \mathbf{MX} + \mathbf{Nb}, \qquad \mathbf{M}, \mathbf{N} \in \mathbb{R}^{n\times n},$$

which means the the limit of the sequence $(\mathbf{X}^k)$ defined by

$$\begin{cases} \mathbf{X}^0 & \text{given guess solution} \\ \mathbf{X}^{k+1} = \Phi(\mathbf{X}^k) \end{cases} \quad k = 0, 1, 2, \ldots. \tag{2.59}$$

The convergence is ensured if $\varrho(\mathbf{M}) < 1$, where $\varrho(\mathbf{M})$ is the spectral radius of the matrix $\mathbf{M}$. The classical affine-linear method splits the matrix $\mathbf{B}$ as

$$\mathbf{B} = \mathbf{L} + \mathbf{D}_1 + \mathbf{U},$$

where $\mathbf{L}$ is the strictly lower triangular matrix, $\mathbf{D}_1$ is the nonsingular diagonal matrix and $\mathbf{U}$ is the strictly upper triangular matrix. The Jacobi method is obtained by choosing

$$\mathbf{M} = -\mathbf{D}_1^{-1}\left(\mathbf{L} + \mathbf{U}\right), \qquad \mathbf{N} = \mathbf{D}_1^{-1},$$

therefore the iteration can be writen as

$$\mathbf{D}_1\left(\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}\right) = -\left(\mathbf{B}\mathbf{X}^{(k)} - \mathbf{b}\right)$$

or

$$\mathbf{X}^{(k+1)} = \mathbf{D}_1^{-1}\left(-\left(\mathbf{L} + \mathbf{U}\right)\mathbf{X}^{(k)} + \mathbf{b}\right).$$

The Gauss-Seidel method is obtained by choosing

$$\mathbf{M} = -\left(\mathbf{D}_1 + \mathbf{L}\right)^{-1}\mathbf{U}, \qquad \mathbf{N} = \left(\mathbf{D}_1 + \mathbf{L}\right)^{-1},$$

It is well known that if the matrix $\mathbf{B}$ is diagonally dominant then both the Jacobi and Gauss-Seidel methods converge [11]. Although the matrices arising in our simulations are diagonally dominant we will not use these methods because they are not efficient.

2.5.2 Krylov subspace methods for linear systems

Krylov subspace methods are among the most powerful methods available for solving large, sparse linear systems. The key points of Krylov subspace methods follow. A initial guess $\mathbf{X}^{(0)}$ is given and used to generate a sequence $\mathbf{X}^{(1)}, \mathbf{X}^{(2)}, \cdots$ for the linear system (2.58) such that

$$\begin{aligned} \mathbf{X}^{(i)} &\in \mathbf{X}^{(0)} + \text{span}\left\{\mathbf{r}^{(0)}, \mathbf{B}\mathbf{r}^{(0)}, \cdots, \mathbf{B}^{i-1}\mathbf{r}^{(0)}\right\} \\ \mathbf{r}^{(0)} &= \mathbf{b} - \mathbf{B}\mathbf{X}^{(0)}, \end{aligned}$$

via the following iterations

$$\mathbf{X}^{(i+1)} = \mathbf{X}^{(i)} + \alpha^{(i)}\mathbf{d}^{(i)}, \;\; i = 1, 2, \cdots$$

where $\alpha^{(i)}$ is a dynamic constant and $\mathbf{d}^{(i)}$ is a search direction. Krylov subspace methods differ by the way to choose or to find $\alpha^{(i)}$ and $\mathbf{d}^{(i)}$. The well known examples are the conjugate gradient method (CG) for symmetric positive definite systems and the minimun residual method (MINRES) for symmetric indefinite systems [11]. For many sparse $n \times n$ matrices CG and MINRES can be performed in $\mathcal{O}(n)$ flops.

In this work we will use for 3D simulations a variant of the iterative Krylov solver, the Bi-Conjugate Gradients Stabilized Method (Bi-CGStab) as implemented in Matlab [36]. Bi-CGStab does not have the restriction on the type of matrix $\mathbf{B}$.

2.6 Péclet number flow and Courant–Friedrichs–Lewy (CFL) number

Usually, standard time integration schemes need to satisfy the CFL condition [37] for stability and convergence. If the upwinding technique is not used for the advection flux, a Péclet number condition [11, 37] is also needed. The local or grid Péclet number $\mathrm{Pe}_{\mathrm{loc}} = \max_i \mathrm{Pe}_i$ where Pe_i is computed over each control volume as

$$\mathrm{Pe}_i := \frac{\max\limits_{\sigma \text{ edge of } i} |q_{i,\sigma}|}{\|\mathbf{D}_i\|_\infty}.$$

$\mathbf{D}_i$ is the mean value of the diffusion matrix over the control volume i and $q_{i,\sigma}$ is the integral of the velocity over the edge σ for the control volume i. If $\mathrm{Pe}_{\mathrm{loc}} > 1$, local transport is dominated by advection and if $\mathrm{Pe}_{\mathrm{loc}} < 1$, local transport is dominated by diffusion.

The grid CFL number is defined as $\nu = \max_i \nu_i$ where

$$\nu_i = \frac{\left(\max\limits_{\sigma \text{ edge of } i} |\overline{q}_{i,\sigma}| \right) \Delta t}{\sup\limits_{(\mathbf{x},\mathbf{y}) \in i^2} |\mathbf{x} - \mathbf{y}|}.$$

$\overline{q}_{i,\sigma}$ is the velocity over the edge σ for the control volume i. The CFL condition requires that $\nu < C$, where C is a constant depending on the particular equation, the space discretization method and often $\mathrm{Pe}_{\mathrm{loc}}$ probably.

Chapter 3

Exponential integrators for advection-dominated reactive transport in anisotropic heterogeneous porous media

In this chapter, we present two exponential time integrators in conjunction with a finite volume discretisation in space for simulating transport by advection and diffusion including chemical reactions in highly heterogeneous porous media representative of geological reservoirs. These numerical integrators are based on the variation of constants solution and solving the linear system exactly. This is at the expense of computing the exponential of the stiff matrix comprising the finite volume discretization. Using real Léja points or a Krylov subspace technique to approximate the exponential makes these methods competitive compared to standard finite difference-based time integrators. We investigate two exponential time integrators, the second-order accurate Exponential Euler Midpoint (EEM) scheme and the Exponential Time Differencing of order one (ETD1). All our numerical examples, which include advection-diffusion-reaction simulations performed on the classical SPE10 test case [1], demonstrate that our methods are highly competitive compared to standard semi-implicit and implicit time integrators. Hence they hence comprise an efficient and accurate method for simulating non-linear advection dominated transport in geological formations. The results of this chapter are presented in our [15, 16, 38].

3.1 Introduction

The ADR equation (2.8) can be discretised in space by the full range of spatial discretisations (e.g, finite differences, finite volumes, or finite elements) and each method comprises its own body of literature. However a fundamental challenge remains. How to integrate in time the system of stiff ODEs, representing transport and reaction processes evolving over multiple time scales, in a stable, accurate and efficient way while avoiding non-physical oscillations (e.g, [39, 40]). The key problem in porous media flow is to overcome the limitations of stability criteria, such as the Courant-Friedrich-Levy condition, when resolving the huge variation in competing transport and reaction rates. Common methods include implicit or adaptive time-stepping (e.g, [41,42]) and operator splitting techniques (e.g, [43,44]). Comparatively new methods are streamline-based simulations where transport is computed along the time-of-flight [45, 46], adaptive mesh refinement to focus the computational effort around the moving fronts and resolve them accurately [47], or event-based simulations where only those regions are only updated where an event (i.e. chemical reaction or transport) occurs [48, 49].

The family of exponential integrators date back to the 1960's (see [8] and [9] for history and detailed references). These methods are based on approximating the corresponding integral formulation of the non-linear part of the differential equation and solving the linear part exactly and computing the exponential of a matrix. Sidje [50] used the Krylov subspace technique and Padé approximation to solve the linear system of ODEs based on variation of constants. Cox and Matthews [14] developed the family of exponential time differencing methods for solving non-linear stiff ODEs. They present the instability issue for computing non-diagonal matrix exponential functions, the so called φ-functions. Kassam and Trefethen [9] used a fourth order exponential time differencing method and the contour integral technique for computing the matrix exponential functions to solve the Kuramoto-Sivashinsky and Allen-Cahn PDEs in one dimension. Berland et al. [51] used a Padé approximation to compute the matrix exponential of φ-functions and provided a package for exponential integrators which is efficient in one dimension.

Although exponential integrators have the advantage that they solve the linear part exactly in time, this is at the price of computing the exponential of a matrix, a notorious problem in numerical analysis [10]. However, new developments in real fast Léja points and Krylov subspace techniques for computing functions of the matrix exponential has revived interest in these methods. The real fast Léja points technique is based on matrix interpolation polynomials at spectral Léja sequences [52, 53]. The Krylov subspace technique is based on the idea of projecting the operator on a "small" Krylov subspace of the matrix via the Arnoldi process [50, 54].

In two and three dimensions, the real fast Léja points technique [53, 55–57] and Krylov subspace technique [55, 56] have been used to implement the matrix exponential of φ-functions efficiently in linear advection diffusion equations. The real fast Léja points technique is also used for the exponential Euler-Midpoint integrator scheme for solving non-linear ADRs [13] and for the exponential Rosenbrock-type integrators for solving semi-linear parabolic PDEs [58]. Simulations have been carried out for homogeneous media with constant dispersion tensors, uniform velocity fields, and low Péclet number flows using finite difference methods or finite element methods for spatial discretisations. In contrast to previous work, we consider anisotropic and heterogeneous media, the exponential time differencing method of order one and Exponential Euler Midpoint with the finite volume discretisation in space and examine high Péclet number flows.

The aim of this chapter is to give the time and space convergence proof of the ETD1 scheme, which is a new result for the finite volume method, compare the performance of the ETD1 and EEM schemes in terms of efficiency and accuracy to standard semi-implicit and fully implicit schemes for the solution of non-linear ADRs in highly heterogeneous porous media with largely varying Péclet number flows. That is situations where transport is locally dominated either by diffusion or advection. We use 2D and 3D simulations and finite volume discretisations to demonstrate the efficiency and the accuracy of the exponential schemes ETD1 and EEM. In the implementation of the ETD1 scheme we also compare the efficiency of the real Léja fast points technique with the Krylov technique for computing the matrix exponential while EEM is implemented only with the Krylov subspace technique.

3.2 Exponential integrators for ADR

3.2.1 Finite volume space discretization and discrete mild solution

Consider our model ADR problem given by (2.8) with the corresponding presure- velocity system given in (2.2). As we describe in Chapter 2, for a given admissible mesh $\mathcal{T}$ in the sense of Definition 2.7 for (2.8) and (2.2), the pressure–velocity system (2.2) is solved to obtain the integral of the velocity using in the semi discrete version of (2.8) given in (2.53) by: Find $X_h(t) \in V_h$ such that

$$\begin{cases} \dfrac{dX_h}{dt} + A_h X_h = P_h R(X_h) \qquad t \in (0,T] \\ \\ X_h(0) = X_{0h} \end{cases} \tag{3.1}$$

with the corresponding discrete variational form

$$\begin{cases} \langle \dfrac{d}{dt} X_h, \varphi \rangle_{0,h} + a_h(X_h(t), \varphi) = \langle R(X_h(t), \varphi \rangle_{0,h}, \quad \forall \varphi \in V_h, \quad t \in (0,T]\,, \\ \\ X_h(0) = X_{h\,0}, \end{cases} \tag{3.2}$$

where the bilinear form a_h is given in (2.49), $\langle -, - \rangle_{0,h}$ is defined in (2.50) and h denotes the maximum mesh size of the admissible mesh $\mathcal{T}$.

Following the definition of the discrete $H^1(\Omega)$ seminorm and the discrete $H_0^1(\Omega)$ norm in [35], we have the following definition of the discrete $\mathbb{H}$ norm. The space $\mathbb{H}$ is defined in (2.32).

Definition 3.1 ***[Discrete*** $\mathbb{H}$ ***norm]***

Let $\mathcal{T}$ be an admissible finite volume mesh in the sense of Definition 2.7. Let $X(\mathcal{T})$ be the space of the functions constant in each control volume of $\mathcal{T}$. For $u \in X(\mathcal{T})$, the $\mathbb{H}$ norm of u is defined by

$$\|u\|_{1,\mathcal{T}} = \left(\sum_{\sigma \in \mathcal{E}_{int} \cup \partial\Omega_D} \tau'_\sigma \left(D_\sigma u \right)^2 \right)^{1/2} \tag{3.3}$$

where

$$\begin{aligned} \tau'_\sigma &= \frac{mes(\sigma)}{d_\sigma} \\ D_\sigma u &= |u_i - u_j| \qquad if \qquad \sigma = i|j \in \mathcal{E}_{int} \\ D_\sigma u &= |u_i| \qquad\quad if \qquad \sigma \in \partial\Omega_D. \end{aligned}$$

During our analysis, to simplify the expression of the discrete bilinear form a_h, we assume without loss of generality that $\partial\Omega = \partial\Omega_D$, which implies that $\mathbb{H} = H_0^1(\Omega)$. We also assume that the porosity ϕ is constant.

As in [35] we make the following assumption.

Assumption 3.2 *[**Regularity of the admissible mesh** $\mathcal{T}$]*
We assume that the restriction of $\mathbf{D}$ *to any* $B_i \in \mathcal{T}$ *belongs to* $C^1(B_i, \mathbb{R}^{d\times d})$, $q_j \in C^1\overline{\Omega})$ *and that there exists* $\zeta_1 > 0$ *and* $\zeta_2 > 0$ *such that*

$$\begin{cases} \zeta_1 h^2 \le \text{mes}(B_i) \le \zeta_2 h^2, \qquad \forall B_i \in \mathcal{T}, \qquad \forall i \in \mathcal{I} \\ \zeta_1 h \le \text{mes}(\sigma) \le \zeta_2 h, \qquad \forall \sigma \in \mathcal{E} \\ \zeta_1 h \le d_\sigma \le \zeta_2 h, \qquad \forall \sigma \in \mathcal{E}. \end{cases} \tag{3.4}$$

Assumption 3.2 allows the following V_h- ellipticity of a_h.

Theorem 3.3 *Under the regularity of the admissible mesh* $\mathcal{T}$ *in Assumption 3.2, there exists a constant* $\alpha > 0$ *such that*

$$a_h(v_h, v_h) \ge \alpha \, \|v_h\|^2_{1,\mathcal{T}} \qquad \forall v_h \in V_h. \tag{3.5}$$

Proof. The proof is the same as the one in [11, Theorem 6.15, page 275] by using (3.38)-(3.39) for the diffusion flux. ■

The following V_h- ellipticity of a_h implies that $-A_h$ is a sectorial on $L^2(\Omega)$ (uniformly in h) i.e. there exists C_1, $\theta \in (\frac{1}{2}\pi, \pi)$, such that

$$\|(\lambda I + A_h)^{-1}\|_{L(L^2(\Omega))} \le \frac{C}{|\lambda|}, \qquad\qquad \lambda \in S_\theta, \tag{3.6}$$

where $S_\theta = \big\{\lambda \in \mathbb{C} : \lambda = \rho e^{i\phi},\ \rho > 0,\ 0 \le |\phi| \le \theta\big\}$.

The discrete operator $-A_h$ therefore is the infinitesimal generator of bounded analytic semigroup (or exponential operator) $S_h(t) := e^{-t\,A_h}$ on V_h such that

$$S_h(t) := e^{-t\,A_h} = \frac{1}{2\pi i}\int_{\mathcal{C}} e^{t\,\lambda}(\lambda I + A_h)^{-1} d\lambda\,, \qquad t > 0 \tag{3.7}$$

where $\mathcal{C}$ denotes a path that surrounds the spectrum of $-A_h$. In the sequel we will use both notations $S_h(t)$ and $e^{-t\,A_h}$ for the analytic semigroup (or exponential operator) generated by $-A_h$. The notation $S_h(t)$ will be used for the proofs of convergence and $e^{-t\,A_h}$ for computation.

As for the continuous case, Duhamel's principle implies that the solution of (2.53) is represented by the following integral equations (mild form)

$$X_h(t) = S_h(t)X_{0h} + \int_0^t S_h(t-s)P_hR(X_h(s))ds, \qquad t \in [0,T]\,. \tag{3.8}$$

The linearity and continuity of P_h defined in (2.54) with the Lipschitz condition of R in (2.29) ensure that P_hR satisfies the following Lipschitz condition

$$\begin{aligned} \|P_hR(u) - P_hR(v)\|_{H^{-1}(\Omega)} &\leq C\|R(u) - R(v)\|_{H^{-1}(\Omega)} \\ &\leq C(\mathcal{B})\|u - v\|,\ \forall u, v \in \mathcal{B} \subset V. \end{aligned} \tag{3.9}$$

where $\mathcal{B}$ is a bounded set. As in Theorem 2.5, the unique mild solution X_h of (3.1) is ensured.

3.2.2 Time discretization and Numerical schemes

We first introduce the Exponential Time Differencing stepping scheme of order one (ETD1) for the ADR problem (2.8). For simplicity we consider a constant time-step $\Delta t > 0$.

At time $t_m = m\Delta t \in [0,T]$, the mild solution (3.8) is given by

$$X_h(t_m) = S_h(t_m)X_{0h} + \int_0^{t_m} S_h(t_m - s)P_hR(X_h(s))ds. \tag{3.10}$$

Then, given the exact solution at the time t_m, we can construct the corresponding solution at t_{m+1} as

$$X_h(t_{m+1}) = S_h(\Delta t)X_h(t_m) + \int_0^{\Delta t} S_h(\Delta t - s)P_hR(X_h(t_m + s))ds. \tag{3.11}$$

Note that the expression in (3.11) is still an exact solution. The idea behind exponential time differencing is to approximate $P_hR(X_h(t_m+s))$ by a suitable polynomial [9, 14]. We consider the simplest case where $P_hR(X_h(t_m+s))$ is approximated by the constant $P_hR(X_h(t_m))$ and the corresponding scheme (ETD1) is given by

$$X_h^{n+1} = e^{-\Delta t A_h} X_h^n + \Delta t \varphi_1(-\Delta t A_h) P_h R(X_h^m, t_m) \tag{3.12}$$

where

$$\varphi_1(-\Delta t A_h) = (-\Delta t\, A_h)^{-1} \left(e^{-\Delta t A_h} - I\right) = \frac{1}{\Delta t}\int_0^{\Delta t} e^{-(\Delta t - s)A_h} ds.$$

Note that the ETD1 scheme in (3.12) can be rewritten as

$$X_h^{m+1} = X_h^m + \Delta t \varphi_1(-\Delta t A_h)(-A_h X_h^m + P_h R(X_h^m)). \tag{3.13}$$

This new expression has the advantage that it is computationally more efficient as only one matrix exponential function needs to be evaluated at each step.

To obtain the second order accurate method with the function φ_1 as in [13], we first linearize locally in time the stiff ODE system (3.1) as follows:

$$\begin{aligned} \frac{dY}{dt} &= -A_h X_h^m + P_h R(X_h^m, t)) + J(X_h^m, t)(Y - X_h^m), \qquad (3.14)\\ Y(t_m) &= X_h^m \\ J(X,t) &= -A_h + \partial_X P_h R(X,t), \qquad t \in [t_m, t_{m+1}]. \end{aligned}$$

Applying the midpoint exponential rule [59, 60] to (3.14) gives the approximation $Y^{m+1} = X_h^{m+1}$ of $X(t_{m+1})$ by

$$\begin{cases} X_h^{m+1} = X_h^m + \Delta t\, \varphi_1\left(\Delta t\; J(X_h^m, t_{m+1/2})\right)\left(-A_h X_h^m + P_h R(X_h^m, t_{m+1/2})\right) \\ \\ J(X_h^m, t_{m+1/2}) = -A_h + \partial_X P_h R(X_h^m, t_{m+1/2}) \\ \\ t_{m+1/2} = (t_{m+1} + t_m)/2 = (2m+1)\Delta t/2. \end{cases} \tag{3.15}$$

The scheme will be called Exponential Euler–Midpoint scheme (EEM). To understand the EEM scheme, it is important to notice that for a general linear ODEs

$$\frac{dy}{dt} = B(t)y + g(t), \qquad y(0) = y_0 \tag{3.16}$$

the midpoint exponential rule scheme is given by

$$y^{m+1} = y^m + \Delta t \varphi_1(\Delta t B(t_{m+1/2})) \left(B(t_{m+1/2}) y^m + g(t_{m+1/2}) \right). \tag{3.17}$$

The scheme EEM is second order like the midpoint exponential rule [13, 59, 60]. It is important to notice that for linear advection–diffusion problems, the schemes EEM and ETD1 are the same and give the exact solution in time up to the tolerance used in the evaluation of the exponential function φ_1. In the sequel we will analyze the time and space convergence of ETD1 since our goal is to extend the ETD1 scheme to stochastic problems.

3.3 Convergence analysis of the ETD1 scheme

We assume that the unique mild solution X of ADR problem (2.8) is the classical solution of (2.8) i.e. X is twice continuously differentiable with respect to $\mathbf{x}$ and differentiable with respect to t. We also assume that R is twice differentiable respect to X and $\mathbf{x}$.

Theorem 3.4 *Consider the mild solution X of the ADR (2.8) and the numerical solution (3.13) given by the ETD1 scheme. Let $\mathcal{B} \subset V$ be bounded, let $t^* = t^*(\mathcal{B})$ defined in Theorem 2.5. Assume that $X(t_m) \in \mathcal{B}$ and $X_h^m \in \mathcal{B}$ with $t_m = m\Delta t \leq T \leq t^*$ and that the finite volume mesh $\mathcal{T}$ satisfies Assumption 3.2. Set $X_{0h} = P_h X_0$, assume that $X_0 \in \mathcal{D}(A)$ and $\|P_h X_0 - X_0\| \leq Ch$, then the following estimates hold:*

If the reaction term R satisfies the Lipschitz condition (2.15), then (2.29) hold and

$$\|X(t_m) - X_h^m\|_{0,h} \leq C(\mathcal{B}) \left(\Delta t^{1-\epsilon} + h \right)$$

for $\epsilon \in (0, 1/2)$ small enough.

If the reaction term R satisfies the local Lipschitz condition from $(L^2(\Omega), \|.\|)$ to $(L^2(\Omega), \|.\|)$ i.e.

$$\|R(u) - R(v)\| \leq C(\mathcal{B}) \|u - v\|, \qquad \forall u, v \in \mathcal{B}, \tag{3.18}$$

then

$$\|X(t_m) - X_h^m\|_{0,h} \leq C(\mathcal{B}) \left(\Delta t + h \right),$$

where $C(\mathcal{B}) = C(\mathcal{B}, \Omega, X, R, \mathbf{D}, \mathbf{q}, T, \zeta_1, \zeta_2)$.

The proof follows in Section 3.3.2, but we need some preparatory results first.

3.3.1 Preparatory results

Proposition 3.5 *Let P_h be the finite volume projection defined in (2.54). Then there exists a positive constant C_0 such that the following inequality holds*

$$\|A_h^{-1/2} P_h f\| \leq C_0 \|f\|_{H^{-1}(\Omega)}, \qquad f \in C(\Omega). \tag{3.19}$$

Proof. The proof is the same as in [32, page 12] for the finite element method. ■

Proposition 3.6 [***Interpolation error***]

Let $\mathcal{T}$ be an admissible mesh in the sense of Definition 2.7 and $\mathcal{T}_h$ its dual Delaunay triangulation ($\{x_i\}$ are vertices of $\mathcal{T}_h$) . Let $I_h : C(\overline{\Omega}) \to V_h$ defined by

$$I_h(u) = \sum_{i \in \mathcal{T}} u(\mathbf{x}_i) \varphi_{\mathbf{x}_i}, \qquad u \in C(\overline{\Omega}) \tag{3.20}$$

where $\{\varphi_{\mathbf{x}_i}\}_{i \in \mathcal{T}}$ is the nodal basis corresponding to $\{\mathbf{x}_i\}_{i \in \mathcal{T}}$ in the sense of finite element method ($\varphi_{x_i}(x_j) = \delta_{i,j}$). If $u \in C^2(\overline{\Omega})$, then there exists a positive constant $C_0(u)$ such that the following estimate holds

$$\|u - I_h(u)\| \leq C_0(u) h^2. \tag{3.21}$$

If $u \in C([0,T], C^2(\overline{\Omega}))$, then

$$\|u(t) - I_h(u(t))\| \leq C_0(u,T) h^2, \qquad \forall t \in [0,T]. \tag{3.22}$$

Proof. *See [11, Section 3.4, Theorem 3.29, page 139 or Exercise 3.25 page 147] or [17, Theorem 17.1, page 132].* ■

Lemma 3.7 *Let X be the mild solution of (2.8) given in (2.36). Let $\mathcal{B} \subset V$ be a bounded set and $t_1, t_2 \in [0,T] \subset [0, t^*(\mathcal{B})]$, $t_1 < t_2$. The following estimates hold :*

- *(i) If $X_0 \in \mathcal{D}(A)$ then*

$$\|X(t_2) - X(t_1)\| \;\leq\; C(t_2 - t_1)^{1-\epsilon} \left(\|X_0\|_2 + \sup_{0 \leq s \leq T} \|R(X(s))\| \right),$$

  for $\epsilon \in (0, 1/2)$ small enough.

- *(ii) If $X_0 \in \mathcal{D}(A)$ and R satisfies the local Lipschitz condition in (3.18) then*

$$\|X(t_2) - X(t_1)\| \;\leq\; C(\mathcal{B})(t_2 - t_1) \left(\|X_0\|_2 + \sup_{0 \leq s \leq T} \|R(X(s))\| \right).$$

Proof. Part (i).

Consider the difference

$$\begin{aligned} & X(t_2) - X(t_1) \\ &= (S(t_2) - S(t_1))\, X_0 + \left(\int_0^{t_2} S(t_2 - s) R(X(s)) ds - \int_0^{t_1} S(t_1 - s) R(X(s)) ds \right) \\ &= I + II, \end{aligned} \tag{3.23}$$

so that $\|X(t_2) - X(t_1)\| \leq \|I\| + \|II\|$. We estimate each of the terms $\|I\|$ and $\|II\|$. For $\|I\|$, using Prposition 2.6 yields

$$\|I\| = \|S(t_1)A^{-1}(\mathrm{I} - S(t_2 - t_1))A^1 X_0\| \leq C(t_2 - t_1)\|X_0\|_2.$$

For the term II, we have

$$\begin{aligned} II &= \int_0^{t_1} (S(t_2 - s) - S(t_1 - s)) R(X(s)) ds + \int_{t_1}^{t_2} S(t_2 - s) R(X(s)) ds \\ &= II_1 + II_2, \end{aligned}$$

with

$$\|II\| \leq \|II_1\| + \|II_2\|.$$

We now estimate each term $\|II_1\|$ and $\|II_2\|$. For $\|II_1\|$

$$\begin{aligned} \|II_1\| &= \| \int_0^{t_1} (S(t_2 - s) - S(t_1 - s)) R(X(s)) ds \| \\ &\leq \int_0^{t_1} \|(S(t_2 - s) - S(t_1 - s)) R(X(s))\| ds \\ &\leq \left(\int_0^{t_1} \|(S(t_2 - s) - S(t_1 - s))\|_{L(L^2(\Omega))} ds \right) \left(\sup_{0 \leq s \leq T} \|R(X(s))\| \right). \end{aligned}$$

For $\epsilon \in (0, 1/2)$ small enough, using Proposition 2.6 yields

$$\begin{aligned}
\|II_1\| &\leq \left(\int_0^{t_1} \|S(t_1-s)A^{1-\epsilon}A^{-1+\epsilon}(\mathrm{I}-S(t_2-t_1))\|_{L(L^2(\Omega))}ds\right)\left(\sup_{0\leq s\leq T}\|R(X(s))\|\right)\\
&\leq \left(\int_0^{t_1} \|A^{1-\epsilon}S(t_1-s)A^{-1+\epsilon}(\mathrm{I}-S(t_2-t_1))\|_{L(L^2(\Omega))}ds\right)\left(\sup_{0\leq s\leq T}\|R(X(s))\|\right)\\
&\leq C(t_2-t_1)^{1-\epsilon}\left(\int_0^{t_1}(t_1-s)^{-1+\epsilon}ds\right)\left(\sup_{0\leq s\leq T}\|R(X(s))\|\right)\\
&\leq C(t_2-t_1)^{1-\epsilon}\left(\sup_{0\leq s\leq T}\|R(X(s))\|\right).
\end{aligned}$$

For $\|II_2\|$, using the fact that the semigroup is bounded, we have

$$\begin{aligned}
\|II\| &= \|\int_{t_1}^{t_2} S(t_2-s)R(X(s))ds\|\\
&\leq \left(\int_{t_1}^{t_2}\|S(t_2-s)R(X(s))\|ds\right)\\
&\leq \left(\int_{t_1}^{t_2}\|R(X(s))\|ds\right)\\
&\leq C(t_2-t_1)\left(\sup_{0\leq s\leq T}\|R(X(s))\|\right).
\end{aligned}$$

Hence

$$\|II\| \leq \|II_1\| + \|II_2\|) \leq C(t_2-t_1)^{1-\epsilon}\left(\sup_{0\leq s\leq T}(\|R(X(s))\|)\right).$$

Combining previous estimations of $\|I\|$ and $\|II\|$ ends the proof of part (i).

Proof of part (ii) We consider again the difference in (3.23). The difference with the proof of part (i) comes from the estimation of II_1. This time we write

$$\begin{aligned}
II_1 &= \int_0^{t_1}(S(t_2-s)-S(t_1-s))R(X(s))ds\\
&= \int_0^{t_1}(S(t_2-s)-S(t_1-s))\left(R(X(s))-R(X(t_1))\right)ds\\
&+ \int_0^{t_1}(S(t_2-s)-S(t_1-s))R(X(t_1))ds\\
&= II_{11}+II_{12}.
\end{aligned}$$

If R satisfies the local Lipschitz condition from $(L^2(\Omega), \|.\|)$ to $(L^2(\Omega), \|.\|)$ given in (3.18), then using the result in part (i) together with Proposition 2.6 yields

$$\begin{aligned}
\|II_{11}\| &\leq \left(\int_0^{t_1} \|S(t_2-s)-S(t_1-s)\|_{L(L^2(\Omega))}\|R(X(s))-R(X(t_1)\|ds\right) \\
&\leq C(\mathcal{B})\left(\int_0^{t_1} \|S(t_2-s)-S(t_1-s)\|_{L(L^2(\Omega))}\|X(s)-X(t_1)\|ds\right) \\
&\leq C(\mathcal{B})\left((t_2-t_1)\int_0^{t_1}(t_1-s)^{-\epsilon}ds\right) \\
&\leq C(\mathcal{B})\,(t_2-t_1)\,,
\end{aligned}$$

for $\epsilon \in (0, 1/2)$ small enough. We also have

$$\begin{aligned}
\|II_{12}\| &\leq \|R(X(t_1)\|\|\int_0^{t_1}(S(t_2-s)-S(t_1-s))ds\|_{L(L^2(\Omega))} \\
&\leq C(\mathcal{B})\|\int_0^{t_1} S(t_2-s)-S(t_1-s)ds\|_{L(L^2(\Omega))}.
\end{aligned}$$

Using the two transformations $y = t_2 - s,\ y = t_1 - s$ we find we find

$$\begin{aligned}
\|II_{12}\| &= C(\mathcal{B})\|\int_{t_2-t_1}^{t_2} S(y)dy - \int_0^{t_1} S(y)dy\|_{L(L^2(\Omega))} \\
&= C(\mathcal{B})\|\int_{t_2-t_1}^{t_1} S(s)ds + \int_{t_1}^{t_2} S(y)dy - \int_0^{t_1} S(y)dy\|_{L(L^2(\Omega))} \\
&= C(\mathcal{B})\|\int_{t_1}^{t_2} S(y)dy - \int_0^{t_2-t_1} S(y)dy\|_{L(L^2(\Omega))} \\
&\leq C(\mathcal{B})(t_2-t_1).
\end{aligned}$$

The estimate of II_1 ends the proof. ■

Lemma 3.8 *[Gronwall lemma [61]]*
Let $u(t)$ and $g(t)$ be nonnegative continuous functions on $I = [0, \infty)$ for which the inequality

$$u(t) \leq M + \int_0^t g(s)u(s)ds, \qquad t \in I$$

holds, where M is a nonnegative constant. Then

$$u(t) \leq M \exp\left[\int_0^t g(s)ds\right], \qquad t \in I \tag{3.24}$$

Lemma 3.9 *[Discrete Gronwall lemma [62]]*

Let sequence $t_n = n\Delta t \leqslant T$. If the sequence of nonnegative numbers ϵ_n satisfies the inequality

$$\epsilon_n \leqslant a\ \Delta t \sum_{j=1}^{n-1} t_{n-j}^{-\beta} \epsilon_j + b\ t_n^{-\sigma} \tag{3.25}$$

for $0 \leqslant \beta, \sigma < 1$ and $a, b \geqslant 0$, then the following estimate holds:

$$\epsilon_n \leqslant C\ b\ t^{-\sigma} \tag{3.26}$$

where the constant C depends on β, σ, a, T.

3.3.2 Proof of Theorem 3.4

In the followng proof, C_i and C_i', $i = 1, \cdots, 6$ are positive constants.

Proof. We use the equivalence of the norms $\|.\|$ and $\|.\|_{0,h}$ in V_h as the mesh $\mathcal{T}$ is regular [11].

Using the triangle inequality yields

$$\begin{aligned} \|X(t_m) - X_h^m\|_{0,h} &\leq \|X(t_m) - X_h(t_m)\|_{0,h} + \|X_h(t_m) - X_h^m\|_{0,h} \\ &= I + II. \end{aligned} \tag{3.27}$$

Let us estimate I. Integrating equation (2.8) over each control volume $i \in \mathcal{T}$ and using the divergence theorem yields

$$\int_i X_t(\mathbf{x}, t) d\mathbf{x} - \sum_{\sigma \in \mathcal{E}_i} \int_\sigma (\mathbf{D}\nabla X - \mathbf{q}X) \cdot \mathbf{n}_\sigma d\sigma = \int_i R(\mathbf{x}, t, X(\mathbf{x}, t)) d\mathbf{x}. \tag{3.28}$$

For $t \in [0, T]$, $i \in \mathcal{T}$ and $\sigma \in \mathcal{E}_i$ using the same notation as in [35], let us set

$$\begin{cases} R_{i,\sigma}(t) = \dfrac{1}{\text{mes}(\sigma)} \left[\dfrac{\text{mes}(\sigma)\mu_\sigma}{d(i,j)} (X_i(t) - X_j(t)) + \int_\sigma \mathbf{D}\nabla X \cdot \mathbf{n}_\sigma d\sigma \right], \\ r_{i,\sigma}(t) = \dfrac{1}{\text{mes}(\sigma)} \left[q_{i,\sigma} X_{i,+} - \int_\sigma \mathbf{q}X(t) \cdot \mathbf{n}_\sigma \right], \\ \rho_i(t) = X(\mathbf{x}_i, t) - \dfrac{1}{\text{mes}(i)} \int_i X(\mathbf{x}, t) d\mathbf{x}, \\ \varrho_i(t) = R(\mathbf{x}_i, t, X_i(t)) - \dfrac{1}{\text{mes}(i)} \int_i R(\mathbf{x}, t, X(\mathbf{x}, t)) d\mathbf{x}. \end{cases} \tag{3.29}$$

Assuming that the unique solution X of (2.8) is the regular, Taylor expansion yields

$$\begin{cases} X_t(\mathbf{x},t) = X_t(\mathbf{x}_i,t) + s_i(\mathbf{x},t), & |s_i(\mathbf{x},t)| \leq C_1\left(X,T\right)h \\ \\ \int_i X_t(\mathbf{x},t)d\mathbf{x} = \mathrm{mes}(i)X_t(\mathbf{x}_i,t) + S_i, \quad S_i = \int_i s_i(\mathbf{x},t)d\mathbf{x}, & |S_i| \leq \mathrm{mes}(i)C_1\left(X,T\right)h. \end{cases} \tag{3.30}$$

Using similar results to in [35] for general elliptic operators, and the regularity of X with respect to t in the compact set $[0,T]$ yields

$$\begin{cases} |R_{i,\sigma}(t)| \leq C_2\left(\mathbf{D},X,T\right)h, \\ \\ |r_{i,\sigma}(t)| \leq C_2'\left(\mathbf{q},X,T\right)h, \\ \\ |R_{i,\sigma}(t)| + |r_{i,\sigma}(t)| \leq C_3\left(\mathbf{q},\mathbf{D},X,T\right)h, \\ \\ |\rho_i(t)| \leq C_3'\left(X,T\right)h. \end{cases} \tag{3.31}$$

Using the fact that R is twice differentiable with respect to X and $\mathbf{x}$, we also have

$$\begin{aligned} &\mathrm{mes}(i)\varrho_i(t) \\ &= \mathrm{mes}(i)R(\mathbf{x}_i,t,X_i(t)) - \int_i R(\mathbf{x},t,X(\mathbf{x},t))d\mathbf{x} \\ &= \int_i \left(R(\mathbf{x}_i,t,X_i(t)) - R(\mathbf{x},t,X(\mathbf{x},t))\right)d\mathbf{x}, \\ &= \mathrm{mes}(i)\left(R(\mathbf{x}_i,t,X_i(t)) - R(\mathbf{x}_i,t,X(\mathbf{x}_i,t)) - \frac{\partial R}{\partial X}(\mathbf{x}_i,t,X(\mathbf{x}_i,t))(X_i(t) - X(\mathbf{x}_i,t))\right) \\ &\quad +\kappa(\mathbf{x}_i,t,X,R). \end{aligned}$$

Using the Lipschitz condition (2.15) and the fact that

$$|\kappa(\mathbf{x}_i,t,X,R)| \leq \mathrm{mes}(i)C_4(R,T,X,B)\,h,$$

yields

$$\mathrm{mes}(i)\varrho_i(t) \leq \mathrm{mes}(i)\left(C_4'(B,\Omega,R,T,X)|X_i(t) - X(\mathbf{x}_i,t)| + C_4(R,T,X)\,h\right). \tag{3.32}$$

Subtracting the first equation of (2.47) from (3.28) and using previous expressions yields

$$\begin{cases} \text{mes}(i)\dfrac{de_i(t)}{dt} + \sum\limits_{\sigma\in\mathcal{E}_i} G_{i,\sigma}(t) + W_{i,\sigma}(t) + c_0\text{mes}(i)e_i(t) \\ \qquad\qquad = \int_i \left(R(\mathbf{x}_i,t,X_i(t)) - R(\mathbf{x},t,X(\mathbf{x},t))\right) d\mathbf{x} \\ \\ \qquad\qquad +c_0\text{mes}(i)\rho_i(t) - \sum\limits_{\sigma\in\mathcal{E}_i}\text{mes}(\sigma)(R_{i,\sigma}+r_{i,\sigma}) - S_i(t), \quad \forall i\in\mathcal{T} \end{cases} \tag{3.33}$$

with

$$\begin{cases} e_i(t) = X(\mathbf{x}_i,t) - X_i(t), \qquad t\in[0,T]\,, \\ \\ G_{i,\sigma}(t) = -\tau_\sigma(e_j(t)-e_i(t)), \qquad \sigma = i|j, \\ \\ G_{i,\sigma}(t) = \tau_\sigma e_i(t), \quad i\subset\partial\Omega, \\ \\ W_{i,\sigma}(t) = q_{i,\sigma}(X(\mathbf{x}_{\sigma,+},t) - X_{\sigma,+}(t)), \end{cases} \tag{3.34}$$

and

$$\begin{cases} \mathbf{x}_{\sigma,+} = \begin{cases} \mathbf{x}_i \ \text{ if } \mathbf{q}\cdot\mathbf{n}_\sigma \geq 0, \\ \qquad\qquad\qquad \sigma = i|j, \\ \mathbf{x}_j \ \text{ if } \ \mathbf{q}\cdot\mathbf{n}_\sigma < 0, \end{cases} \\ \\ \mathbf{x}_{\sigma,+} = \begin{cases} \mathbf{x}_i \ \text{ if } \ \ \mathbf{q}\cdot\mathbf{n}_\sigma \geq 0, \\ \qquad\qquad\qquad\qquad \sigma\in\mathcal{E}_i\cap\partial\Omega. \\ \mathbf{x}_\sigma, \quad \mathbf{x}_\sigma\in\partial\Omega \ \text{ if } \ \mathbf{q}\cdot\mathbf{n}_\sigma < 0, \end{cases} \end{cases} \tag{3.35}$$

Multipling equation (3.33) by $e_i(t)$ and summing for $i\in\mathcal{T}$ yields

$$\begin{cases} \sum\limits_{i\in\mathcal{T}}\left[\dfrac{\text{mes}(i)}{2}\dfrac{d(e_i^2(t))}{dt} + \sum\limits_{\sigma\in\mathcal{E}_i} e_i(t)(G_{i,\sigma}(t)+W_{i,\sigma}(t)) + c_0\,\text{mes}(i)e_i^2(t)\right] \\ \\ = \sum\limits_{i\in\mathcal{T}} e_i(t)\left[\int_i \left(R(\mathbf{x}_i,t,X_i(t)) - R(\mathbf{x},t,X(\mathbf{x},t))\right) d\mathbf{x}\right] \\ \\ +\sum\limits_{i\in\mathcal{T}}\left[c_0\,\text{mes}(i)\rho_i(t)e_i(t) - \sum\limits_{\sigma\in\mathcal{E}_i}\text{mes}(\sigma)e_i(t)(R_{i,\sigma}(t)+r_{i,\sigma}(t)) - e_i(t)S_i(t)\right]. \end{cases} \tag{3.36}$$

Let $e_{\mathcal{T}}(t)$ a piecewise constant function defined by

$$e_{\mathcal{T}}(t) = e_i(t), \qquad \text{for} \quad \mathbf{x} \in i, \quad i \in \mathcal{T}, \quad t \in [0,T]\,. \tag{3.37}$$

Assumption 3.2 for the regularity of $\mathcal{T}$ and the fact that the coefficients of the diffusion tensor $\mathbf{D}$ are bounded implies that there exists two constants $C_5(\Omega,\zeta_1,\zeta_2,\mathbf{D})$ and $C_5'(\Omega,\zeta_1,\zeta_2,\mathbf{D})$ such that

$$C_5 \le \mu_\sigma = d_{i,j}\frac{D_{i,\sigma}D_{i,\sigma}}{D_{i,\sigma}\,d_{i,\sigma} + D_{i,\sigma}d_{i,\sigma}} \le C_5', \qquad \sigma = i|j, \tag{3.38}$$

and

$$C_5 \le \mu_\sigma = D_{i,\sigma} \le\ C_5', \qquad \sigma \in \mathcal{E}_i \cap \partial\Omega, \tag{3.39}$$

so that

$$C_5 \le \mu_\sigma\ \le\ C_5'\,, \qquad \forall \sigma \in \mathcal{E}, \tag{3.40}$$

where μ_σ is defined in (2.43) and (2.42).

Using the fact that the transmissibility given in (2.41) is symmetric, i.e. $\tau_{i|j} = \tau_{j|i}$ and reorganizing the summation yields

$$\begin{cases} C_5\|e_{\mathcal{T}}(t)\|_{1,\mathcal{T}}^2 \le \|e_{\mathcal{T}}(t)\|_{1,h}^2 \le C_5'\,\|e_{\mathcal{T}}(t)\|_{1,\mathcal{T}}^2 \\ \\ \|e_{\mathcal{T}}(t)\|_{1,\mathcal{T}}^2 := \displaystyle\sum_{\sigma\in\mathcal{E}} |D_\sigma e_{\mathcal{T}}(t)|^2 \frac{\mathrm{mes}(\sigma)}{d_\sigma} \quad \text{(discrete norm of } H_0^1(\Omega) \text{ in Definition 3.1)} \\ \\ \|e_{\mathcal{T}}(t)\|_{1,h}^2 := \displaystyle\sum_{i\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_i} e_i(t)G_{i,\sigma}(t), \end{cases} \tag{3.41}$$

where

$$\begin{cases} |D_\sigma e_{\mathcal{T}}(t)| = |e_i(t) - e_j(t)|, \qquad \text{if} \qquad \sigma = i|j, \\ \\ |D_\sigma e_{\mathcal{T}}(t)| = |e_i(t)|, \qquad \text{if} \qquad \sigma \in \mathcal{E}_i \cap \partial\Omega. \end{cases} \tag{3.42}$$

Setting $e_{\sigma,+}(t) = X(\mathbf{x}_{\sigma,+},t) - X_{\sigma,+}(t)$, as in [35] for stationary elliptic advection–diffusion–reaction and using the fact that $\nabla\cdot\mathbf{q} = 0$, yields

$$\begin{aligned} \sum_{i\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_i} e_i(t)W_{i,\sigma}(t) &= \sum_{i\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_i} q_{i,\sigma}e_i(t)(X(\mathbf{x}_{\sigma,+},t) - X_{\sigma,+}(t)) \\ &= \sum_{i\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_i} q_{i,\sigma}e_i(t)e_{\sigma,+}(t) \ge 0. \end{aligned} \tag{3.43}$$

Using (3.43) in the expression (3.36) yields

$$\begin{cases} \frac{1}{2}\sum_{i\in\mathcal{T}}\text{mes}(i)\frac{d(e_i^2(t))}{dt} + \|e_{\mathcal{T}}(t)\|_{1,h}^2 + c_0\|e_{\mathcal{T}}(t)\|_{0,h}^2 \leqslant C_4'\|e_{\mathcal{T}}(t)\|_{0,h}^2 + C_4\, h\sum_{i\in\mathcal{T}}\text{mes}(i)|e_i(t)| \\ \\ +c_0C_3'\, h\sum_{i\in\mathcal{T}}\text{mes}(i)|e_i(t)| + \sum_{i\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_i}\text{mes}(\sigma)e_i(t)(R_{i,\sigma}(t)+r_{i,\sigma}(t)) + C_1\, h\sum_{i\in\mathcal{T}}\text{mes}(i)|e_i(t)|. \end{cases} \tag{3.44}$$

The continuity of the diffusion and advection flux at each interface yields

$$R_{i,\sigma}(t) = -R_{j,\sigma}(t), \qquad r_{i,\sigma}(t) = -r_{j,\sigma}(t), \qquad \text{for } \sigma = i|j \in \mathcal{E}_{int}.$$

Set

$$R_\sigma(t) = |R_{i,\sigma}(t)|, \;\; r_\sigma(t) = |R_{i,\sigma}(t)|, \qquad i \in \mathcal{T}, \qquad \sigma \in \mathcal{E}_{int}.$$

Using the relation (3.31), the Cauchy-Schwarz inequality as in [35] for stationary elliptic problems and reordering the summation over the edges yields

$$\begin{aligned} &\sum_{i\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_i}\text{mes}(\sigma)e_i(t)(R_{i,\sigma}(t)+r_{i,\sigma}(t)) \\ &\leqslant \sum_{\sigma\in\mathcal{E}} D_\sigma e_{\mathcal{T}}(t)(R_\sigma(t)+r_\sigma(t)) \\ &\leqslant \left(\sum_{\sigma\in\mathcal{E}}\frac{\text{mes}(\sigma)}{d_\sigma}(D_\sigma e_{\mathcal{T}}(t))^2\right)^{\frac{1}{2}}\left(\sum_{\sigma\in\mathcal{E}}\text{mes}(\sigma)d_\sigma(R_\sigma+r_\sigma)^2\right)^{\frac{1}{2}}. \end{aligned}$$

Using the fact that $\sum_{\sigma\in\mathcal{E}}\text{mes}(\sigma)d_\sigma \leqslant d\,\text{mes}(\Omega))$ and relation (3.41) yields

$$\begin{aligned} \sum_{i\in\mathcal{T}}\sum_{\sigma\in\mathcal{E}_i}\text{mes}(\sigma)e_i(t)(R_{i,\sigma}(t)+r_{i,\sigma}(t)) &\leqslant C_3\, h\,(\text{mes}(\Omega)\, d)^{\frac{1}{2}}\|e_{\mathcal{T}}(t)\|_{1,\mathcal{T}} \\ &\leqslant (C_5)^{-1}C_3\, h\,(\text{mes}(\Omega)\, d)^{\frac{1}{2}}\|e_{\mathcal{T}}(t)\|_{1,h}. \end{aligned} \tag{3.45}$$

For any constant $C > 0$, Young's inequality yields

$$\begin{cases} |C\, h\sum_{i\in\mathcal{T}}\text{mes}(i)e_i(t)| = |\sum_{i\in\mathcal{T}}(C\ h\ \text{mes}(i)^{\frac{1}{2}})(\text{mes}(i)^{\frac{1}{2}}e_i(t))| \leqslant \frac{1}{2}\|e_{\mathcal{T}}(t)\|_{0,h}^2 + \frac{1}{2}C^2h^2\ \text{mes}(\Omega) \\ \\ C\, h\|e_{\mathcal{T}}(t)\|_{1,h} \leqslant \frac{1}{2}C^2h^2 + \frac{1}{2}\|e_{\mathcal{T}}(t)\|_{1,h}^2. \end{cases} \tag{3.46}$$

Using expression (3.45) and (3.46) in expression (3.44) yields

$$\begin{cases} \frac{1}{2}\left[\sum_{i\in\mathcal{T}}\text{mes}(i)\frac{d(e_i^2(t))}{dt} + \|e_{\mathcal{T}}(t)\|_{1,h}^2 + c_0\,\|e_{\mathcal{T}}(t)\|_{0,h}^2\right] \leqslant (C_4'/2+1)\|e_{\mathcal{T}}(t)\|_{0,h}^2 + C_6\, h^2 \\ \\ C_6 = C_6(C_1, C_3, C_3', C_4, C_5). \end{cases} \tag{3.47}$$

Bounding the left hand side of expression (3.47) below yields

$$\sum_{i\in\mathcal{T}}\text{mes}(i)\frac{d(e_i^2(s))}{dt} \leqslant (C_4'+2)\|e_{\mathcal{T}}(s)\|_{0,h}^2 + 2C_6\,h^2, \qquad \forall s\in[0,T]. \tag{3.48}$$

Integrating both side of expression (3.48) through interval $[0,t]\,,\ 0\leq t\leq T$ yields

$$\|e_{\mathcal{T}}(t)\|_{0,h}^2 \leq \|e_{\mathcal{T}}(0)\|_{0,h}^2 + 2\,C_6\,T\,h^2 + (C_4'+2)\int_0^t \|e_{\mathcal{T}}(s)\|_{0,h}^2 ds, \qquad \forall t\in[0,T]. \tag{3.49}$$

Applying the discrete Gronwall Lemma 3.9 yields

$$\|e_{\mathcal{T}}(t)\|_{0,h}^2 \leq C\left(\|e_{\mathcal{T}}(0)\|_{0,h}^2 + h^2\right), \tag{3.50}$$

with $C = C(\mathcal{B},\Omega,X,R,\mathbf{D},\mathbf{q},T,\zeta_1,\zeta_2)$.

Then

$$I = \|X(t_m) - X_h(t_m)\|_{0,h} = \|e_{\mathcal{T}}(t_m)\|_{0,h} \leq C\left(\|X_0 - X_{0h}\|_{0,h} + h\right). \tag{3.51}$$

If $X_{0h} = P_hX_0$, as we have assumed

$$\|X_0 - X_{0h}\|_{0,h} = \|X_0 - P_hX_0\|_{0,h} \leq Ch, \tag{3.52}$$

which is true for more meshes (see [11]), we therefore have

$$I = \|X(t_m) - X_h(t_m)\|_{0,h} = \|e_{\mathcal{T}}(t_m)\|_{0,h} \leq Ch. \tag{3.53}$$

Let us estimate II. From (3.10) and (3.12) we have

$$X_h(t_m) = S_h(t_m)X_{0h} + \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_h(t_m-s)P_hR(s,X_h(s))ds, \tag{3.54}$$

and

$$X_h^m = S_h(t_m)X_{0h} + \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_h(t_m-s)P_hR(t_k,X_h^k)ds. \tag{3.55}$$

The smoothing properties of the semigroup S_h in Proposition 2.6 together with Proposition 3.5 and the equivalence $\|.\|\equiv\|.\|_{0,h}$ in V_h yields

$$\begin{aligned}
&\|X_h(t_m) - X_h^m\|_{0,h}\\
&\equiv \|X_h(t_m) - X_h^m\|\\
&= \|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} A_h^{1/2}S_h(t_m-s)A_h^{-1/2}P_h\left(R(s,X_h(s)) - R(t_k,X_h^k)\right)ds\|\\
&\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m-s)^{-1/2}\|A_h^{-1/2}P_h\left(R(s,X_h(s)) - R(t_k,X_h^k)\right)\|ds,
\end{aligned}$$

and

$$\begin{aligned}
\|X_h(t_m) - X_h^m\|_{0,h} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} \| \left(R(s, X_h(s)) - R(t_k, X_h^k) \right) \|_{H^{-1}(\Omega)} ds \\
&\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} \| R(s, X_h(s)) - R(s, X(s)) \|_{H^{-1}(\Omega)} ds \\
&\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} \| R(s, X(s)) - R(t_k, X_h^k) \|_{H^{-1}(\Omega)} ds \\
&= II_1 + II_2. \qquad (3.56)
\end{aligned}$$

For $s \in [0, T]$, using Proposition 3.6 yields

$$\begin{aligned}
\|X_h(s) - X(s)\| &\leq \|X_h(s) - I_h(X(s)) + I_h(X(s)) - X(s)\| \\
&\leq (\|X_h(s) - I_h(X(s))\| + \|I_h(X(s)) - X(s)\|) \\
&\leq \left(\|X_h(s) - I_h(X(s))\| + C(X, T) h^2 \right). \qquad (3.57)
\end{aligned}$$

Since $X_h(s) - I_h(X(s)) \in V_h$, the equivalence $\|.\| \equiv \|.\|_{0,h}$ and the estimate of the term I yields

$$\begin{aligned}
\|X_h(s) - I_h(X(s))\| &\leq C \|X_h(s) - I_h(X(s))\|_{0,h} \\
&= C \|X_h(t_m) - X_h^m\|_{0,h} \\
&\leq C(\mathcal{B}, \Omega, X, R, \mathbf{D}, \mathbf{q}, \zeta_1, \zeta_2) h \qquad (3.58)
\end{aligned}$$

Using the Lipschitz condition (2.29) with (3.58) and (3.57) yields

$$II_1 \leq C(\mathcal{B}, \Omega, X, R, \mathbf{D}, \mathbf{q}, T, \zeta_1, \zeta_2) h. \qquad (3.59)$$

We also have

$$\begin{aligned}
II_2 &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} \| R(s, X(s)) - R(t_k, X(t_k)) \|_{H^{-1}(\Omega)} ds \\
&+ \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} \| R(t_k, X(t_k)) - R(t_k, X_h^k) \|_{H^{-1}(\Omega)} ds \\
&= II_2^1 + II_2^2. \qquad (3.60)
\end{aligned}$$

Using Lemma 3.7 and the Lipschitz condition (2.29) yields

$$\begin{aligned} II_2^1 &\leq C(\mathcal{B})\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-s)^{-1/2}\|X(s)-X(t_k)\|_{0,h}ds \\ &\leq C(\mathcal{B})\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-s)^{-1/2}(s-t_k)^{1-\epsilon}ds \\ &\leq C(\mathcal{B})\Delta t^{1-\epsilon}\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-s)^{-1/2}ds \\ &\leq C(\mathcal{B})\Delta t^{1-\epsilon}\int_0^{t_m}(t_m-s)^{-1/2}ds \leq C(\mathcal{B},T)\Delta t^{1-\epsilon}, \end{aligned} \tag{3.61}$$

with $\epsilon \in (0,1/2)$ small enough.

Using Lipschitz condition (2.29) and Proposition 3.6

$$II_2^2 \leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-s)^{-1/2}\|X(t_k)-I_h(X(t_k))+I_h(X(t_k))-X_h^k\|ds \tag{3.62}$$

$$\leq C(X,T)\left(h^2+\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-s)^{-1/2}\|X(t_k)-X_h^k\|_{0,h}ds\right) \tag{3.63}$$

Then

$$II \leq C\left(\left(\Delta t^{1-\epsilon}+h\right)+\int_{t_k}^{t_{k+1}}(t_m-s)^{-1/2}\|X(t_k)-X_h^k\|_{0,h}ds\right),$$

Combining estimates I and II yields

$$\begin{aligned} &\|X(t_m)-X_h^m)\|_{0,h} \\ &\leq C(\mathcal{B})\left(\Delta t^{1-\epsilon}+h+\int_{t_k}^{t_{k+1}}(t_m-s)^{-1/2}\|X(t_k)-X_h^k\|_{0,h}ds\right). \end{aligned} \tag{3.64}$$

Applying the discrete Gronwall Lemma 3.9 in (3.64) ends the first part of the proof.

The second part follows the same approach as the first part by using the second claim of Lemma 3.7. ■

Remark 3.10 *It is important to notice that when the diffusion tensor is diagonal i.e.* $\mathbf{D} = D(\mathbf{x})\mathbf{I}_2$*, the admissible mesh in Definition 2.7 is the usual Voronoi mesh. As a consequence we can drop (3.4) in Assumption 3.2 and the proof will be independent of* ζ_1 *and* ζ_2*.*

3.4 Implementation of the exponential integrator schemes

The key element in the exponential integrators schemes is computing the matrix exponential functions, the so called φ_i-functions. For our schemes ETD1 and EEM we need to implement the φ_1-function and later in this thesis we will also need the φ_0-function (the exponential function).

3.4.1 Padé approximation for φ_i-functions

The family of φ_i-functions used in the exponential integrator schemes are defined for a non-singular matrix z by

$$\begin{cases} \varphi_0(z) = e^z \\ \\ \varphi_i(z) = \left(\varphi_{i-1}(z) - \frac{1}{(i-1)!}\right) z^{-1} \end{cases} \qquad i = 1, 2, \cdots . \tag{3.65}$$

For a given positive integer p, the idea behind the (p, p) Padé approximation of φ_i-functions is to approximate φ_i by a rational fraction such that

$$\varphi_i(z) \approx P^i_{pp}(z) = N^i_p(z) \left(D^i_p(z)\right)^{-1}, \tag{3.66}$$

where the unique polynomials N^i_p and D^i_p are

$$\begin{cases} N^i_p(z) = \frac{p!}{(2p+i)!} \sum_{k=0}^{p} \left[\sum_{l=0}^{k} \frac{(2p+i-l)!(-1)^l}{(l!(p-l)!(i+k-l)!)} \right] z^k \\ \\ D^i_p(z) = \frac{p!}{(2p+i)!} \sum_{k=0}^{p} \frac{(2p+i-k)!}{(k!(p-i)!)} (-z)^k. \end{cases} \tag{3.67}$$

Without approximation we have [51]

$$\varphi_i(z) = N^i_p(z) \left(D^i_p(z)\right)^{-1} + \mathcal{O}(z^{2p+1}). \tag{3.68}$$

The approximation (3.66) is accurate only in the neighborhood of the null matrix, so during implementation we need to scale the matrix to make it small. The implementation of $\varphi_i, i = 0, 1$ is as follows.

- Let s be the smallest integer such that $2^s \geq \|z\|_\infty$.

- Set $z_{scaled} = \dfrac{z}{2^\eta}$, $\eta = \max(0, s+1)$.
- Take the approximation $\varphi_i(z_{scaled}) \approx P^i_{pp}(z_{scaled})$ using the (p,p) Padé approximant from (3.66).
- Undo scaling by using the following relations

$$\begin{cases} \varphi_0(z) \approx \left(P^0_{pp}(2^{-\eta}z)\right)^{2^\eta} = \left(P^0_{pp}(z_{scaled})\right)^{2^\eta} \\ \\ \varphi_1(2x) = \varphi_1(x)\left(\frac{x\varphi_1(x)}{2} + \mathbf{I}\right) \approx P^1_{pp}(x)\left(\frac{xP^1_{pp}(x)}{2} + \mathbf{I}\right), \qquad \mathbf{I} = \text{(identity matrix)}. \end{cases} \tag{3.69}$$

In practice the values of p usually used are $p = 6$ and $p = 7$, giving very good approximations (see [10, 50, 51]).

In the previous implementation if $\eta \gg 1$ the undo scaling stage is time consuming. For large matrices we also need to compute and save inverses of matrices during the simulation, which are very expensive tasks in time and storage. We can then see that implementing the ETD1 and EEM schemes with the Padé approximation for φ_i is not efficient for realistic problems in 2 and 3 dimensions, and it is well known that a standard Padé approximation for a matrix exponential functions is not an efficient method for large scale problems [10,50,51]. Here we focus on the real fast Léja points and the Krylov subspace techniques to evaluate the action of the exponential matrix function $\varphi_i(-\Delta t\, A_h)$ on a vector $\mathbf{v}$, instead of computing the full exponential function $\varphi_i(-\Delta t\, A_h)$ as in a standard Padé approximation. The details of the real fast Léja points technique [50,54] while for the Krylov subspace technique details are given in [52, 53, 57]. We give a brief summary below.

Notice that for the EEM scheme we need to compute at the k^{th} step the action of $\varphi_i\left(-\Delta t\,(A_h + \partial_X P_h R(X^k_h, t_{k+1/2}))\right)$ in the same way as the action of $\varphi_1(-\Delta t\, A_h)$.

3.4.2 Real fast Léja points technique for the action φ_i, $i = 0, 1$

For a given vector $\mathbf{v}$, real fast Léja points approximate $\varphi_i(-\Delta t\, A_h)\mathbf{v}$ by $P_m(-\Delta t\, A_h)\mathbf{v}$, where P_m is polynomial of degree m an interpolating φ_i at the sequence of points $\{\xi_i\}_{i=0}^m$ called spectral real fast Léja points. These points $\{\xi_i\}_{i=0}^m$ belong to the spectral focal interval $[\alpha, \beta]$ of the matrix $-\Delta t A_h$, i.e. the focal interval of the smallest ellipse containing all the eigenvalues of $-\Delta t\, A_h$. This spectral interval can be estimated by the well known

Gershgorin circle theorem [63]. It has been shown that as the degree of the polynomial increases and hence the number of Léja points increases, convergence is achieved [57], i.e.

$$\lim_{m\to\infty} |\varphi_i(-\Delta t A_h)\mathbf{v} - P_m(-\Delta t\, A_h)\mathbf{v}| = 0, \tag{3.70}$$

where $|.|_2$ is the standard Euclidian norm. For a real interval $[\alpha, \beta]$, a sequence of real fast Léja points $\{\xi_i\}_{i=0}^m$ is defined recursively as follows. Given an initial point ξ_0, usually $\xi_0 = \beta$, the sequence of fast Léja points is generated by

$$\prod_{k=0}^{j-1}|\xi_j - \xi_k| = \max_{\xi\in[\alpha,\beta]}\prod_{k=0}^{j-1} \mid \xi - \xi_k \mid \quad j = 1, 2, 3, \cdots. \tag{3.71}$$

We use the Newton's form of the interpolating polynomial P_m given by

$$P_m(z) = \varphi_i\left[\xi_0\right] + \sum_{j=1}^{m}\varphi_i\left[\xi_0, \xi_1, \cdots, \xi_j\right]\prod_{k=0}^{j-1}(z - \xi_k) \tag{3.72}$$

where the divided differences $\varphi_i[\bullet]$ are defined recursively by

$$\begin{cases} \varphi_i\left[\xi_j\right] = \varphi_i(\xi_j) \\ \\ \varphi_i\left[\xi_j, \xi_{j+1}, \cdots, \xi_k\right] := \dfrac{\varphi_i\left[\xi_{j+1}, \xi_{j+2}, \cdots, \xi_k\right] - \varphi_i\left[\xi_j, \xi_{j+1}, \cdots, \xi_{k-1}\right]}{\xi_k - \xi_j}. \end{cases} \tag{3.73}$$

We summarise in Algorithm 1 the steps for computing $\varphi_i(-\Delta t\, A_h)\mathbf{v}$ in the standard way, i.e. by computing (3.73) directly. In our implementation we estimate the focal interval for $-A_h$ only once and precompute a sufficiently large number z of Léja points using the efficient algorithm of Baglama et al. [52] for a focal interval of $-\Delta t A_h$. The data are passed as input parameters during each call of the algorithm and scaled by Δt. Using this approach, we observed the same convergence problems as described by Caliari et al. [57], that is problems arising from round-off errors during the computation of the divided differences (3.73) and from the large capacity of the spectral focal interval $[\alpha, \beta]$. We were able to resolve this issue by reducing the time-step size or by using an algorithm that we will present shortly for minimising rounding errors from the divided differences [64,65] when computing (3.73). Note that although it is advised in [57] to compute the divided differences in quadruple precision we did not find this necessary.

The standard approach cannot produce accurate divided differences with magnitude smaller than machine precision. In [66] it is shown that Léja points for the interval $[-2, 2]$

Algorithm 1 : Standard computation of $\varphi_1(-\Delta t\, A_h)\mathbf{v}$ with real fast Léja points. Error e_m is controlled to a prescribed tolerance tol so that $e_m^{\text{Léja}} < tol$.

Input: $A_h, \mathbf{v}, \Delta t, tol, z$ {matrix, vector, time-step, tolerance, number of Léja points to be generated }

$[\alpha, \beta] = \text{getfocal}(A_h)$ {get the focal interval using the Gershgorin circle theorem [63]}

$\xi = \text{getLeja}(\alpha, \beta, z)$ {generate z fast Léja points from (3.71).}

$d_0 = \varphi_1(\xi_0)$.

$\mathbf{w}_0 = \mathbf{v}$, $\mathbf{p}_0 = d_0\mathbf{w}_0$, $m = 0$ {initialisation}

while $e_m^{\text{Léja}} = |d_m| \times |\mathbf{w}_m| > tol$ **do**

  $\mathbf{w}_{m+1} = (-\Delta t A_h - \xi_m \mathbf{I})\mathbf{w}_m$

  $m = m + 1$

  $d_m = \varphi_1(\xi_m)$

  for $i = 1, \cdots, m$ **do**

    $d_m = \dfrac{d_m - d_{i-1}}{\xi_m - \xi_{i-1}}$ {compute the next divided difference d_m}

  end for

  $\mathbf{p}_m = \mathbf{p}_{m-1} + d_m\mathbf{w}_m$

end while

Output: $\mathbf{p}_m$

assure optimal accuracy, thus for the spectral focal interval $[\alpha, \beta]$ of the matrix $-\Delta t A_h$, it is convenient to interpolate, by a change of variables, the function $\varphi_i(c + \gamma\xi)$ of the independent variable $\xi \in [-2, 2]$ with $c = (\alpha + \beta)/2$ and $\gamma = (\beta - \alpha)/4$. It can be shown [65] that the divided differences of a function $f(c + \gamma\xi)$ of the independent variable ξ at the points $\{\xi_i\}_{i=0}^m \subset [-2, 2]$ are the first column of the matrix function $f(\mathbf{L}_m)$, where

$$\mathbf{L}_m = c\mathbf{I}_{m+1} + \gamma\widehat{\mathbf{L}}_m, \quad \widehat{\mathbf{L}}_m = \begin{pmatrix} \xi_0 & & & & & \\ 1 & \xi_1 & & & & \\ & 1 & \ddots & & & \\ & & \ddots & \ddots & & \\ & & & \ddots & \ddots & \\ & & & & 1 & \xi_m \end{pmatrix}.$$

We then conclude that the divided differences of $\varphi_i(c + \gamma\xi)$ of the independent variable $\xi \in [-2, 2]$ at the points $\{\xi_i\}_{i=0}^m \subset [-2, 2]$ is $\varphi_i(\mathbf{L}_m)e_1^{m+1}$ where e_1^{m+1} is the first standard basis vector of $\mathbb{R}^{m+1}$. Taylor expansion of order p with scaling and squaring is used in [64, 65] to compute $\varphi_i(\mathbf{L}_m)e_1^{m+1}$. In practice the real fast Léja points are computed once in the interval $[-2, 2]$ and reused at each time step during the computation of the divided differences. We use the efficient algorithm of Baglama et al. [52] to compute the real fast Léja points in $[-2, 2]$.

3.4.3 Krylov space subspace technique for the action φ_i, $i = 0, 1$

The main idea of the Krylov subspace technique is to approximate the action of the exponential matrix function $\varphi_i(-\Delta t\, A_h)$ on a vector $\mathbf{v}$ by projection onto a small Krylov subspace $K_m = \text{span}\left\{\mathbf{v}, A_h\mathbf{v}, \ldots, A_h^{m-1}\mathbf{v}\right\}$ (see [50]). The approximation is formed using an orthonormal basis of $\mathbf{V}_m = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_m]$ of the Krylov subspace K_m and of its completion $\mathbf{V}_{m+1} = [\mathbf{V}_m, \mathbf{v}_{m+1}]$. The basis is found by Arnoldi iteration [67] which uses stabilised Gram-Schmidt to produce a sequence of vectors that span the Krylov subspace (see Algorithm 2). Let $\mathbf{e}_i^j$ be the i^{th} standard basis vector of $\mathbb{R}^j$. We approximate $\varphi_i(-\Delta t\, A_h)\mathbf{v}$ by

$$\varphi_i(-\Delta t\, A_h)\mathbf{v} \quad \approx \quad |\mathbf{v}|\mathbf{V}_{m+1}\varphi_i(-\Delta t\, \overline{\mathbf{H}}_{m+1})\mathbf{e}_1^{m+1} \tag{3.74}$$

with

$$\overline{\mathbf{H}}_{m+1} = \begin{pmatrix} \mathbf{H}_m & \mathbf{0} \\ 0, \cdots, 0, h_{m+1,m} & 0 \end{pmatrix} \qquad \text{where} \qquad \mathbf{H}_m = \mathbf{V}_m^T A_h \mathbf{V}_m = [h_{i,j}].$$

The coefficient $h_{m+1,m}$ is recovered in the last iteration of Arnoldi's iteration.

Approximation (3.74) comes from the fundamental relation (see [50])

$$A_h \mathbf{V}_m = \mathbf{V}_m \mathbf{H}_m + h_{m+1,m} \mathbf{v}_{m+1} (\mathbf{e}_m^m)^T \tag{3.75}$$

and the fact that for any $\tau > 0$

$$\begin{aligned} & \tau\varphi_i(-\tau A_h)\mathbf{v} \\ = \;& \tau|\mathbf{v}|\mathbf{V}_m\varphi_i(-\tau\mathbf{H}_m)\mathbf{e}_1^m + |\mathbf{v}| \sum_{j=i+1}^{\infty} h_{m+1,m}\tau^j (\mathbf{e}_m^m)^T \varphi_j(-\tau\mathbf{H}_m)\mathbf{e}_1^m A_h^{j-2}\mathbf{v}_{m+1} \\ = \;& \tau|\mathbf{v}|\mathbf{V}_{m+1}\varphi_i(-\tau\overline{\mathbf{H}}_{m+1})\mathbf{e}_1^{m+1} + |\mathbf{v}| \sum_{j=i+2}^{\infty} h_{m+1,m}\tau^j (\mathbf{e}_m^m)^T \varphi_j(-\tau\mathbf{H}_m)\mathbf{e}_1^m A_h^{j-2}\mathbf{v}_{m+1}, \end{aligned}$$

where

$$\begin{aligned} \varphi_i\left(-\tau\overline{\mathbf{H}}_{m+1}\right) &= \begin{pmatrix} \varphi_i(-\tau\mathbf{H}_m) & \mathbf{0} \\ \tau h_{m+1,m}(\mathbf{e}_m^m)^T \varphi_{i+1}(-\tau\mathbf{H}_m) & 1 \end{pmatrix} \\ &= \begin{pmatrix} \varphi_i(-\tau\mathbf{H}_m) & \mathbf{0} \\ \tau h_{m+1,m}\varphi_{i+1}(-\tau\mathbf{H}_m)[m, 1:m] & 1 \end{pmatrix} \end{aligned} \tag{3.76}$$

(see [50] for more details).

Here for any matrix $\mathbf{M}$, we denote by $\mathbf{M}[i, j:k]$ the row vector whose elements are the elements of the i th row of the matrix $\mathbf{M}$ from the j th column to the k th column, and by $\mathbf{M}[j:k, i]$ the column vector whose elements are the elements of the i th column of the matrix $\mathbf{M}$ from the the j th row to the k th row. Equation (3.76) gives the following expression, needed in equation (3.74).

$$\varphi_i(-\tau\overline{\mathbf{H}}_{m+1})\mathbf{e}_1^{m+1} = \begin{pmatrix} \varphi_i(-\tau\mathbf{H}_m)[1:m, 1] \\ \\ \tau h_{m+1,m}\varphi_{i+1}(-\tau\mathbf{H}_m)[m, 1] \end{pmatrix}. \tag{3.77}$$

Let $\mathbf{c} \in \mathbb{R}^m$ and $p \in \mathbb{N}$ and set

$$\widehat{\mathbf{H}}_{m+p} = \begin{pmatrix} \mathbf{H}_m & \mathbf{c} & 0 & \cdots & 0 \\ & 0 & 1 & \ddots & \vdots \\ & & 0 & \ddots & 0 \\ & & & \ddots & 1 \\ 0 & & & & 0 \end{pmatrix} \in \mathbb{R}^{(m+p)\times(m+p)}. \tag{3.78}$$

Then Sidje [50] showed that for any $\tau > 0$

$$\exp\left(-\tau \widehat{\mathbf{H}}_{m+p}\right) \tag{3.79}$$

$$= \begin{pmatrix} \exp\left(-\tau \mathbf{H}_m\right) & \tau\varphi_1\left(-\tau \mathbf{H}_m\right)\mathbf{c} & \tau^2\varphi_2\left(-\tau \mathbf{H}_m\right)\mathbf{c} & \cdots & \tau^p\varphi_p\left(-\tau \mathbf{H}_m\right)\mathbf{c} \\ & 1 & \frac{\tau}{1!} & \cdots & \frac{\tau^{p-1}}{(p-1)!} \\ & & 1 & \ddots & \vdots \\ & & & \ddots & \frac{\tau}{1!} \\ 0 & & & & 1 \end{pmatrix} \tag{3.80}$$

Since m is small, in implementation we just need to compute

$$\exp(-\Delta t\, \widehat{\mathbf{H}}_{m+2}), \tag{3.81}$$

at each time step by using a Padé approximation's technique [50, 51] with $\mathbf{c} = \mathbf{e}_1^m$. We then deduce the values of $\varphi_i\left(-\Delta t\, \overline{\mathbf{H}}_{m+1}\right)\mathbf{e}_1^{m+1}$ using equation (3.77) with elements in (3.81).

In our implementation we use the function `phiv.m` of the package Expokit [50], which allows us to compute the forward ETD1 or EEM solution using the previous solution while controlling the local error at each iteration for a given tolerance. The function `phiv.m` takes the time step Δt, the matrix A_h, the vectors $\mathbf{u}$ and $\mathbf{v}$, the dimension of the Krylov subspace m, and the desired tolerance *tol* as the input and provides $\mathbf{u} + \Delta t\varphi_1(-\Delta t\, A_h)(-A_h\mathbf{u} + \mathbf{v})$ as the output. In this thesis we will also use the function `expv.m` of the package Expokit [50] for the action of φ_0. This method is accurate for a symmetric matrix with negative eigenvalues but can be less efficient on very large non-symmetric matrices [50, 54].

Algorithm 2 : Arnoldi's algorithm

Initialise: $\mathbf{v}_1 = \dfrac{\mathbf{v}}{|\mathbf{v}|}$ {normalisation}
for $j = 1 \cdots m$ **do**
  $\mathbf{w} = A_h \mathbf{v}_j$
  for $i = 1 \cdots j$ **do**
    $h_{i,j} = \mathbf{w}^T \mathbf{v}_i$ {compute inner product to build elements of the matrix $\mathbf{H}_m$}
    $\mathbf{w} = \mathbf{w} - h_{i,j}\mathbf{v}_i$ {Gram–Schmidt process}
  end for
  $h_{j+1,j} = |\mathbf{w}|$
  $\mathbf{v}_{j+1} = \dfrac{\mathbf{w}}{|\mathbf{w}|}$ {normalisation}
end for

3.5 Numerical experiments of ETD1 scheme in 2D

Through all this section the diffusion tensor is taken to be

$$\mathbf{D} = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \tag{3.82}$$

with $D_1 > 0, D_2 > 0$ and the we take the isotropic permeability $\mathbf{k} = k\mathbf{I}_2$.

To analyse the convergence and efficiency of the ETD1 method for solving ADRs, we apply it to a variety of porous media flow problems and compare it to our standard time-stepping methods, the implicit Euler and semi implicit schemes introduced in the previous chapter. We consider the following four problems:

1. A linear ADR without reaction term, a heterogeneous dispersion tensor, and a non-uniform velocity field representing moderate Péclet number flows, for which an analytical solution exists [68]. We will call this problem **Problem 1**.

2. A non-linear ADR in homogeneous media where transport is controlled equally by advection and diffusion (i.e, Péclet number is 1) for which an analytical solution exists [13]. We will call this problem **Problem 2**.

3. A non-linear ADR for a deterministic permeability field representing a highly idealised fractured porous media. Here transport is entirely dominated by advection (high Péclet number flow). We will call this problem **Problem 3**.

4. A non-linear ADR for a stochastically generated permeability field where transport is locally dominated by either advection or diffusion. We will call this problem **Problem 4**.

In the two latter applications we use the classical Langmuir isotherm to model the sorption of the transported species onto the rock surface, i.e.

$$R(X) = \frac{\lambda\beta X}{1 + \lambda X}.$$

The parameter λ is an adsorption constant and β the maximum amount of the solute that can be adsorbed. We take $\lambda = \beta = 1$ in this work.

For the sake of simplicity we assume that the porosity ϕ is constant in all applications. In all cases we take our domain to be rectangular $\Omega = [0, L_1) \times [0, L_2)$ and use both uniform and non-uniform, rectangular meshes. The time has been normalised by the average flow rate and the domain length in the direction of flow such that the mean of the concentration has traveled through the entire domain at $T = 1$. In each application example, the matrix A_h is pentadiagonal. For a grid size $N_x \times N_y$, the corresponding matrix has the size $N_xN_y \times N_xN_y$ with $5 \times N_xN_y - 2 \times N_x - 6$ non-zero elements.

For the pressure, we take the Dirichlet boundary $\Gamma_D^1 = \{0, L_1\} \times [0, L_2]$ and Neumann boundary $\Gamma_N^1 = (0, L_1) \times \{0, L_2\}$ such that

$$p = \begin{cases} 1 & \text{in} \quad \{0\} \times [0, L_2] \\ \\ 0 & \text{in} \quad \{L_1\} \times [0, L_2] \end{cases}$$

$$-\mathbf{k}\,\nabla p(\mathbf{x}, t) \cdot \mathbf{n} = 0 \quad \text{in} \quad \Gamma_N^1.$$

For the concentration X, we take the Dirichlet boundary $\Gamma_D = \{0\} \times [0, L_2]$ and Neumann boundary $\Gamma_N = \{(0, L_1] \times \{0, L_2\}\} \cup \{\{L_1\} \times [0, L_2]\}$ such that

$$X = 1 \quad \text{in} \quad \Gamma_D \times [0, T]$$

$$-(\mathbf{D}\nabla X)(\mathbf{x}, t) \cdot \mathbf{n} = 0 \quad \text{in} \quad \Gamma_N \times [0, T]$$

$$X_0 = 0 \quad \text{in} \quad \Omega \quad \text{(initial solution)}$$

where $\mathbf{n}$ is the unit outward normal vector to Γ_N (or Γ_N^1).

For applications where we do not have an analytic solution we estimate the global error by

$$\|X_h(t) - X_h^{\Delta t}(t)\| \approx 2\|X_h^{\Delta t}(t) - X_h^{\Delta t/2}(t)\|,$$

as the error estimate of the ETD1 scheme is $\mathcal{O}(\Delta t)$ in time. We denoted by $X_h^{\Delta t}(t)$ the approximation of the solution at time t found with time-step Δt. Unless explicitly stated, the tolerance used for Newton's method and the (ETD1) schemes is 10^{-6} and the Krylov space dimension used is $m = 6$. The tests were performed on a standard PC with a 3 GHz processor and 2GB RAM. Our code was implemented in Matlab 7.7. In the legends of all of our graphs we use the following notation

- "Implicit with Newton" denotes results from the implicit Euler with standard Newton method.
- "Implicit with Newton V" denotes results from the implicit Euler with the variant of Newton method.
- "Léja ETD1" denotes results from ETD1 with real fast Léja points for matrix exponential.
- "Krylov ETD1" denotes results from ETD1 using Krylov subspace technique for the matrix exponential.
- "Semi implicit" denotes results from the semi-implicit scheme.

3.5.1 Homogeneous porous media without reaction term (Problem 1)

We use this problem to examine the scaling of the ETD1 method for problems with different numbers of unknowns and analyse the convergence in space by comparing it to an exact solution [68]. Since the ADR does not contain a reaction term, the problem is linear. The domain is defined as $\Omega = [L_0, L_1) \times [L_0, L_1)$, $L_0 = 0.01$, $L_1 = 2$. The initial time is given as $t_0 = 0.01$. This is necessary because the exact solution is not defined at the origin and

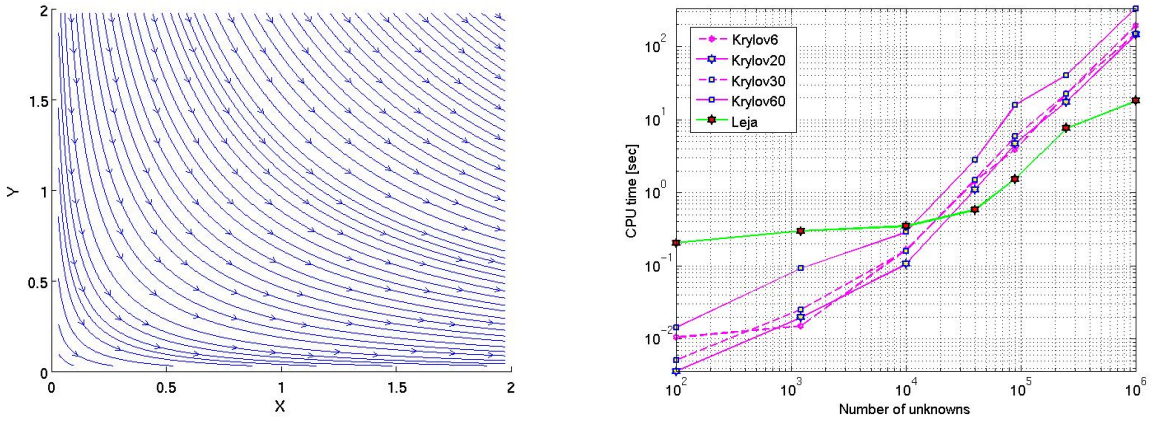
at $t = 0$. The dispersion tensor $\mathbf{D}$ is heterogeneous and its coefficients are given by

$$\begin{cases} D_1(x,y) = D_0 u_0^2 x^2 & (x,y) \quad \in \Omega \\ D_2(x,y) = D_0 u_0^2 y^2 & (x,y) \quad \in \Omega. \end{cases}$$

The velocity field (Fig. 3.1(a)) is given explicitly by

$$\begin{cases} \mathbf{q} = (q_x, q_y)^T \\ q_x(x,y) = u_0 x \quad (x,y) \quad \in \overline{\Omega} \\ q_y(x,y) = -u_0 y \quad (x,y) \quad \in \overline{\Omega} \end{cases} \tag{3.83}$$

where $D_0 = 0.1$ and $u_0 = 2$. The local Péclet number ranges from 21 to 2 as the grid is refined. Initial and boundary conditions are taken according to the exact solution [68], assuming an instantaneous release at a point (x_0, y_0), $x_0 = 1.5$, $y_0 = 1.5$. We take a fixed time-step of $\Delta t = 1/3000$.



(a) (b)

Figure 3.1: Numerical examples for the linear advection-diffusion problem in homogeneous porous media given in [68] (a) shows the streamlines, (b) shows the CPU time as a function of number of unknowns required to evaluate the expression $\varphi_1(-\Delta t\, A_h)(-A_h X_h^0 + P_h R(X_h^0, T_0))$. A standard PC with a 3 GHz processor and 2GB RAM was used for the simulations. The number in the four Krylov curves in (b) denotes the dimension m of the subspace taken.

Figure 3.1(a) shows the streamlines which indicate direction of flow. Figure 3.1(b) shows the CPU time needed to compute single time-step using ETD1 with real Léja points and Krylov techniques as a function of the number of unknowns. The number of the real fast

Léja points used to achieve the given tolerance are 6 for 100 unknowns, and increases to 69 as the grid is refined.

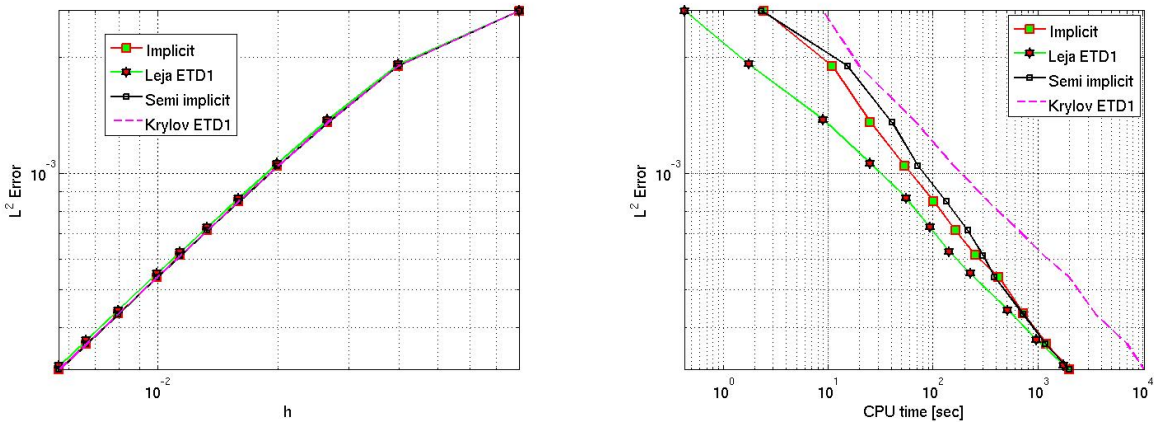
In Figure 3.1(b) we show that good values for the dimension of the Krylov subspace are $m = 20$ and $m = 6$, but $m = 20$ appears to be a slightly better value for this specific example. To our knowledge, there is no rigorous theory that allows us to predict the optimal value for m a priori. For example, the default value used in [50] is $m = 30$ but we observe that this is not the optimal value for our specific example. When m increases, the total number of iterations decreases but a penalty occurs due to the additional time spent in the orthogonalisation process in Algorithm 2 and the corresponding increase in memory requirements. For small m, a penalty can arise from an increase in the number of iterations necessary to achieve a given tolerance, especially if Δt is large, but less time is spent in the orthogonalisation process and the required memory is lower. Since the memory on the PC used in this work is limited to 2 GB, the values of Δt in our application examples are generally small, and we need to compute the action of the matrix exponential function φ_1 on a vector over 3000 times to reach the final time T, we have chosen $m = 6$ as the optimal value for the Krylov subspace dimension in all our applications.

For 10^4 and more unknowns, that is for problem sizes that become representative for real reservoir simulations, the computation of the matrix exponential with real fast Léja points is more efficient than the Krylov technique by a factor of approximately 10, regardless of the Krylov subspace dimension m. Similar results were obtained by [55, 56] for constant dispersion tensor, constant velocity, and low Péclet number flows. Once the matrix size is greater than or equal to 10^4, the CPU time increases linearly with the number of unknowns (Figure 3.1(b)). The time to evaluate a matrix with 10^6 unknowns using 69 real Léja points is 18 seconds. These results suggest that the ETD1 is a scalable solver and is hence probably applicable to large-scale problems with several millions of unknowns that are encountered in 3D reservoir simulations.

Figure 3.2(a) shows a convergence of order $\mathcal{O}(h)$ for the spatial discretisation with fixed time step $\Delta t = 1/3000$. The error in the L^2 norm is computed at time $T = 1$. Figure 3.2(b) shows the L^2 error as a function CPU time, which is depicted in Figure 3.2(a).

The efficiency for solving this linear ADR problem is roughly similar for all methods, that is approximately the same computational cost is required to reduce the numerical

error by a certain increment. Although Figure 3.1(b) indicates that for small number of unknowns the Krylov technique requires significantly less computational effort than the real Léja points method to compute one step with one vector $\mathbf{v}$, Figure 3.2(b) shows that over the course of an entire simulation, which involves many individual time-steps, the local error control reduces this efficiency, therefore Krylov and Léja points methods are comparable. We recall that the Krylov subspace implementation is known to be efficient for symmetric matrices. Here we observe good convergence even for highly nonsymmetric matrices A_h.



(a) (b)

Figure 3.2: (a) Convergence of the L^2 norm at $T = 1$ as a function of the grid size h. (b) The L^2 norm at $T = 1$ as a function of CPU time. Both plots are for the linear ADR in homogeneous porous media without a reaction term (Problem 1) with fixed $\Delta t = 1/3000$. Recall here that the Krylov subspace dimension is fixed to be $m = 6$. The analytical solution exists [68]

.

3.5.2 Homogeneous porous media with a non-linear reaction term (Problem 2)

We now evaluate the ETD1 method for a non-linear ADR problem where the non-linear reaction term is given by $R(X) = -\gamma X^2(1 - X)$. We take $\gamma = 100$, use a constant velocity of $\mathbf{q} = [-0.01, -0.01]^T$, and the dispersion tensor has the entries $D_1 = D_2 = 10^{-4}$. The domain is $\Omega = [0, 1) \times [0, 1)$, which we discretise with $h = \Delta x = \Delta y = 10^{-2}$. The local Péclet number for the flow is 1, that is transport is controlled equally by advection and

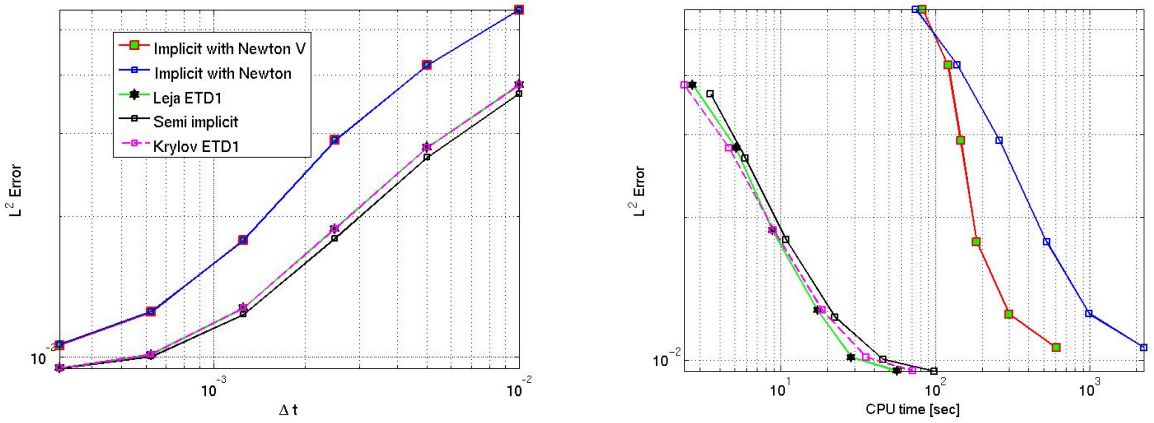
diffusion. The initial condition and boundary conditions are defined with respect to the exact solution [13] given by

$$X(x, y, t) = (1 + \exp(a(x + y - bt) + a(b - 1)))^{-1} \tag{3.84}$$

where $a = \sqrt{\gamma / (4 \times 10^{-4})}$ and $b = -0.02 + \sqrt{\gamma \times 10^{-4}}$.

Figure 3.3(a) shows the convergence as a function of the chosen time-step Δt, measuring the error at the final time $T = 1$. The semi-implicit time-stepping method and the ETD1 methods have similar error constants. All schemes have the same rate of convergence $\mathcal{O}(\Delta t)$.

Figure 3.3(b) shows the L^2 error as a function of CPU time, which is given in Figure 3.3(a). Again, the computational effort to reduce the error by a certain factor is approximately equivalent for both Léja and Krylov subspace techniques. They are also similar to a semi-implicit time integrator. However, all three methods, ETD1 with Léja points and Krylov subspace technique and semi-implicit time-stepping, outperform the implicit time-stepping methods. Those require about 10 times more computational costs to obtain the same numerical error. If the advective component of the flux is included in the non-linear part rather than the linear part for the ETD1 scheme, then the error constant worsens. In this case the graph representing the error would lie between that of the ETD1 or semi-implicit error and implicit error in Figure 3.3(a).



(a) (b)

Figure 3.3: (a) Convergence of the L^2 norm at $T = 1$ as a function of Δt. (b) The L^2 norm at $T = 1$ as a function of CPU time. Both are for the the non-linear ADR in a homogeneous porous medium (Problem 2).

3.5.3 Deterministic heterogeneous porous media and non-linear reaction (Problem 3)

We now test the ETD1 method for a porous media with three parallel high-permeability streaks. This could represent, for example, transport in a highly idealised fracture pattern. The permeability of the three parallel streaks is 100 times greater than the permeability of the surrounding domain (Figure 3.4(a)). Hence flow is diverted from the lower-permeability rocks into the high-permeability matrix (Figure 3.4(b)). The advection rates increase towards the high-permeability streaks and are highest in them. This is clearly visible by the closer spacing of the streamlines in the high–permeability streaks (Figure 3.4(b)).

For the non-linear reaction term we now take the Langmuir sorption isotherm. The domain is given by $\Omega = [0, 2) \times [0, 3)$ and discretised in space with $\Delta x = 3/50$ and $\Delta y = 1/25$. The dispersion tensor is anisotropic with $D_1 = 10^{-3}$, $D_2 = 10^{-4}$. The viscosity is $\mu = 0.1$. The maximum local Péclet number is 2975.4.

Figure 3.4(c) shows the concentration at $t = 0.3$ and Figure 3.4(d) the concentration at $T = 1$. Again, the flow-focusing due to the high-permeability streaks is clearly visible.

Figure 3.5(a) shows the convergence at the final time $T = 1$ in the L^2 norm for varying time-steps Δt. All schemes show convergence rates of $\mathcal{O}(\Delta t)$. There is now a distinct difference between ETD1 method with Krylov or Léja point technique and the implicit and semi-implicit integrators. The ETD1 methods displays a clear improvement in the error constant. Figure 3.5(b) depicts the L^2 error at $T = 1$ as a function of CPU time. The ETD1 based schemes are significantly more accurate and computationally more efficient than (semi-)implicit schemes. They require between 10 and 100 times less computational effort to achieve the same reduction in numerical error. The Léja point method has also a small computational advantage over the Krylov subspace technique.

3.5.4 Stochastic heterogeneous porous media with non-linear reaction (Problem 4)

We finally apply the ETD1 method to a stochastically generated permeability field. Stochastic permeability fields are commonly used to represent the unknown heterogeneity in the subsurface. We use the Karhunen-Loeve numerical expansion [27] to generate the random

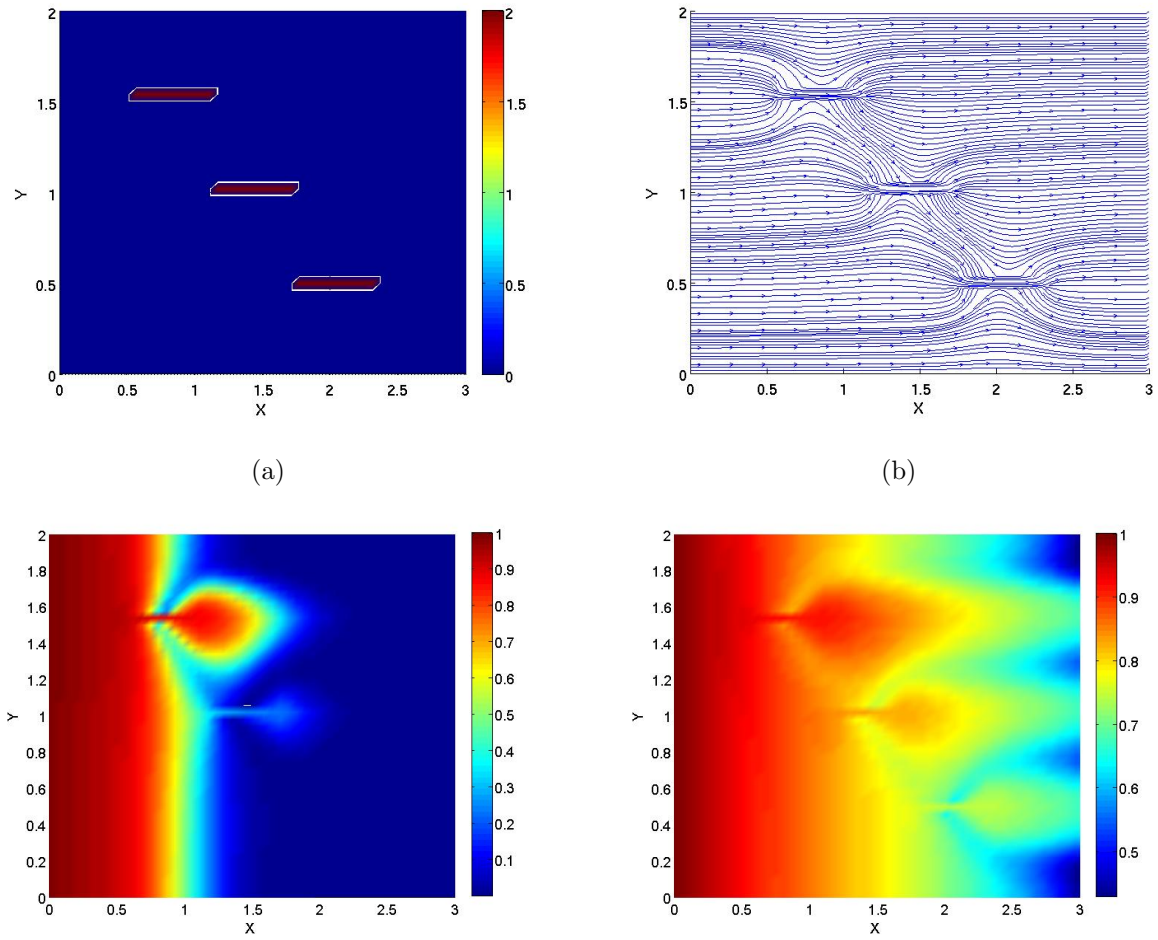
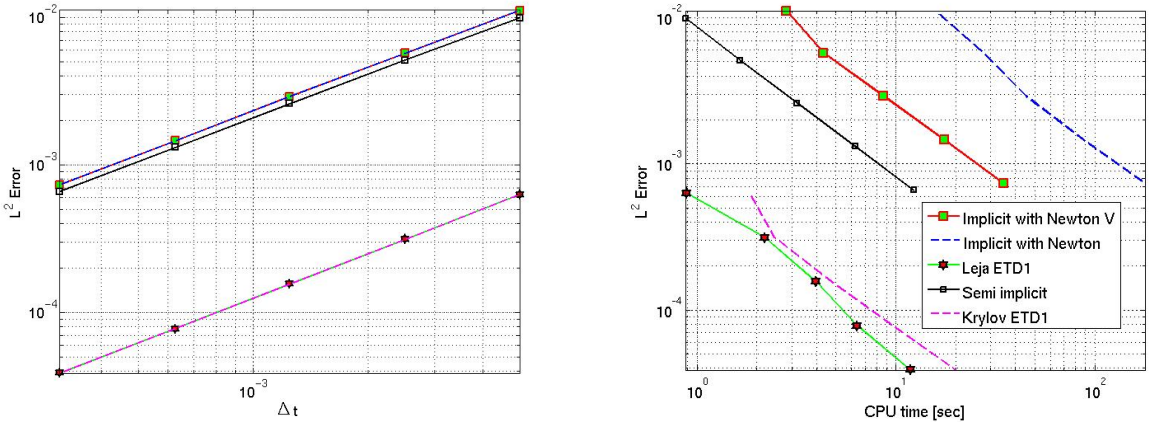


Figure 3.4: Numerical experiments for the non-linear ADR problem in a deterministic heterogeneous porous media (Problem 3). (a) shows the log of permeability field, (b) shows the velocity streamlines, (c) shows the concentration at $t = 0.3$ and (d) shows the concentration field at $T = 1$.

permeability field from a log-normal distribution with an exponentially decaying space correlation. The correlation in the field is given by

$$C_r((x_1, y_1); (x_2, y_2)) = \frac{1}{4b_1b_2} \exp\left(-\frac{\pi}{4}\left[\frac{(x_2 - x_1)^2}{b_1^2} + \frac{(y_2 - y_1)^2}{b_2^2}\right]\right),$$

where b_1 and b_2 are the spatial correlation lengths in x-direction and y-direction, respectively, and given by $b_1 = 0.4$ and $b_2 = 0.2$. We used the first 30 terms in the Kahunen-Loeve expansion (2.7) for the permeability field and used the same field to evaluate all the time integrators. The domain is given by $\Omega = [0, 3) \times [0, 2)$ with $\Delta x = 1/10$ and $\Delta y = 1/15$.



(a) (b)

Figure 3.5: (a) Convergence of the L^2 norm at $T = 1$ as a function of Δt. (b) The L^2 norm at $T = 1$ as a function of CPU time. Both plots are for the non-linear ADR in a deterministic heterogeneous porous medium (Problem 3). Although all time integrators display a convergence rate of $\mathcal{O}(\Delta t)$, there is clear improvement in the error constant. Hence the ETD1 schemes are significantly more efficient than (semi-)implicit methods, with the real Léja point method being most efficient.

The dispersion tensor has the entries $D_1 = 10^{-3}$, $D_2 = 10^{-4}$ and the viscosity $\mu = 1$. The maximum local Péclet number is $Pe_{loc} = 1649.3$.

Figure 3.6(a) shows the log of the permeability field, which varies over 6 orders of magnitude ranging from 10^{-3} to 10^3. Figure 3.6(b) shows the corresponding streamlines, which show how flow is focused into regions of high permeability. Advection rates are significantly higher in regions of high permeability, reflected by the close streamline spacing, compared to regions of low permeability. Figure 3.6(c) shows the concentration at $T = 0.2$ and Figure 3.6(d) the concentration at $T = 1$. Both show flow-focusing into the high permeability regions.

Figure 3.7(a) shows the convergence of the L^2 norm at $T = 1$ as a function of Δt. As in all our previous applications, all schemes have similar convergence rates of $\mathcal{O}(\Delta t)$, but there is a clear improvement in the error constant for the ETD1 schemes. Figure 3.7(b) shows the L^2 error as a function of CPU time. The ETD1 methods clearly outperform the implicit time-integrators, with the Léja points based scheme being slightly more efficient than the Krylov subspace based methods. The latter shows a similar performance to the semi-

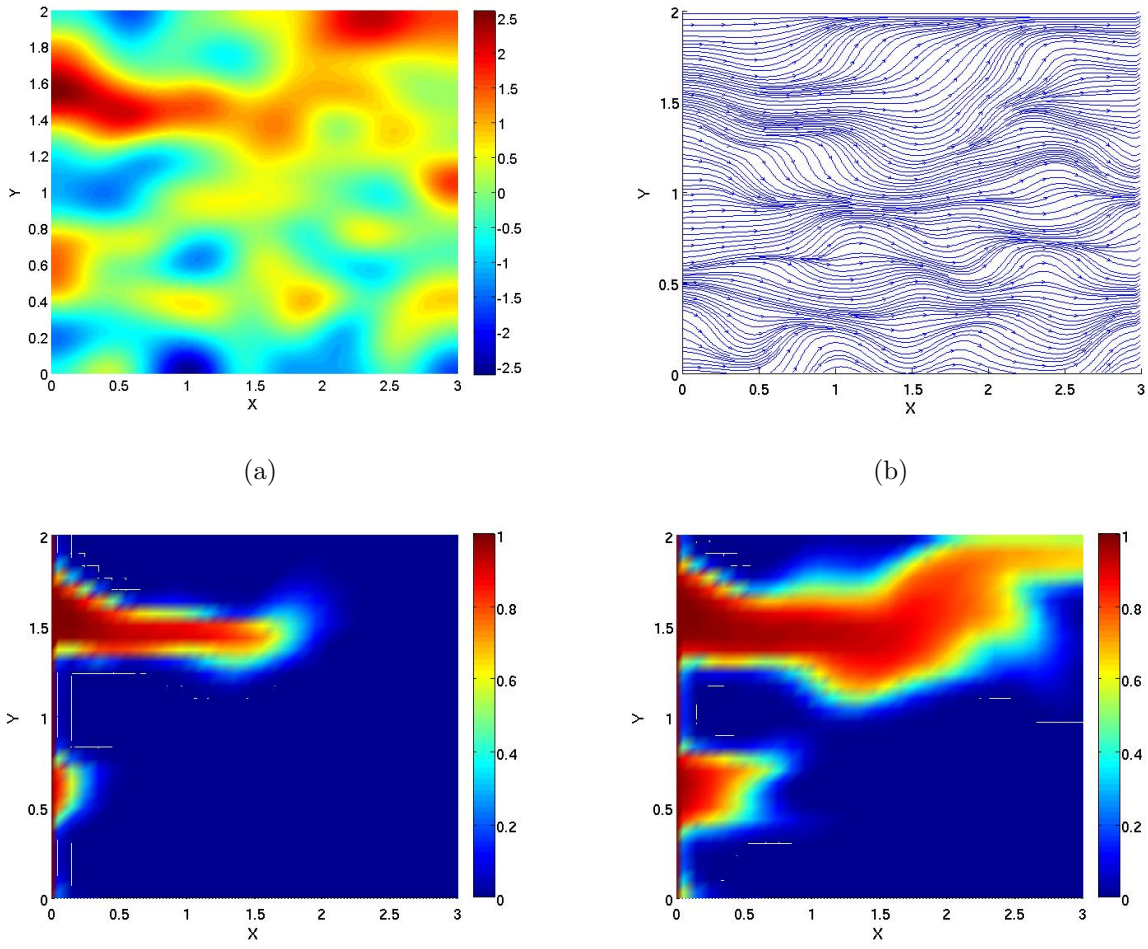
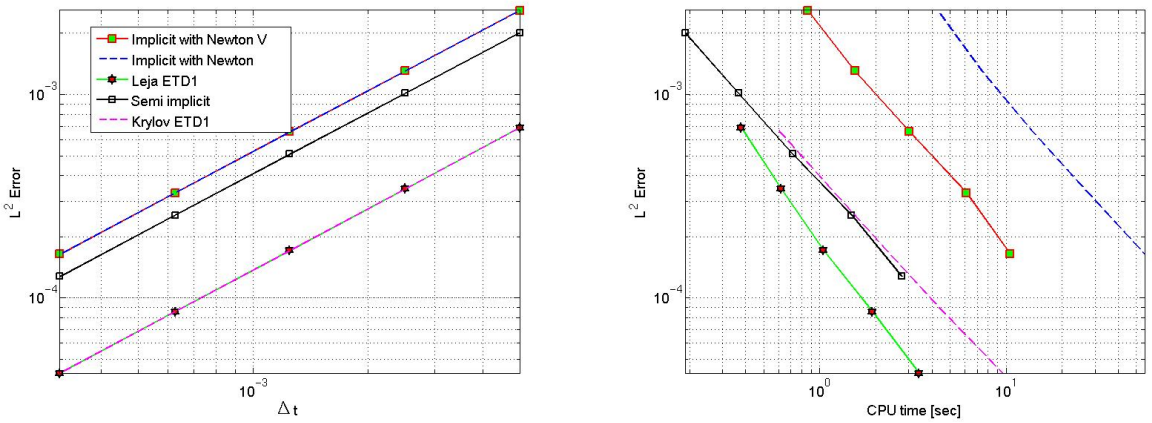


Figure 3.6: Numerical experiments for non-linear ADR problem in a stochastic heterogeneous porous medium (a) shows the log of permeability field, (b) shows the velocity streamlines, (c) shows the concentration at $t = 0.2$ and (d) shows the concentration at $T = 1$.

implicit method. Table 3.1 compares the CPU time necessary to perform 3200 steps of the ETD1 integration using the real Léja points method and the Krylov subspace technique. We analysed how many Léja points are required for the first step for different spatial discretisations ranging from 100×100 grid points to 500×2000 grid points. For the largest problem with 10^6 unknowns only 10 Léja points are required. The total CPU time necessary to find the solution at final time $T = 1$ is 5293.3 seconds.

For the Krylov subspace method (with $m = 6$) the total CPU time required to find the solution at the final time $T = 1$ is 14693 seconds. We observe that the real Léja points



(a) (b)

Figure 3.7: (a) Convergence of the L^2 norm at $T = 1$ as a function of Δt. (b) The L^2 norm at $T = 1$ as a function of CPU time. Both plots are for the non-linear ADR in a stochastically generated porous media (Problem 4).

method seems more efficient than the Krylov implementation, taking approximately half the CPU time. We have tested several values for m and due to the reasons discussed previously, we do not think that the Krylov subspace technique will be more efficient than the real Léja points method for other values of m. Nevertheless, this example demonstrates that ETD1 methods are probably applicable to large-scale 3D reservoir simulations with several million unknowns.

3.6 Numerical experiments of ETD1 and EEM schemes in 3D

In this section, we take Ω to be an open domain of $\mathbb{R}^3$ and solve over a finite time interval $[0, T]$. We take the symmetric dispersion (diffusion) tensor $\mathbf{D}$ to be

$$\mathbf{D} = \begin{pmatrix} D_1 & 0 & 0 \\ 0 & D_2 & 0 \\ 0 & 0 & D_3 \end{pmatrix} \tag{3.85}$$

with $D_1 > 0, D_2 > 0, D_3 > 0$.

N_x	N_y	$M_{\text{Léja}}$	CPU time [sec] Léja	CPU time [sec] Krylov ($m = 6$)
100	100	6	21	54
200	200	7	123	316
100	1000	7	430	889
200	1000	8	911	2349
100	3000	10	1589	2715
500	2000	10	5293	14693

Table 3.1: CPU time for the real Léja points and Krylov subspace methods used in Problem 4 as a function of various grid sizes. N_x is the number of subdivisions in the x direction and N_y the number in the y direction. Table shows the number of Léja points used for the first step $M_{\text{Léja}}$ and the CPU time to perform 3200 steps of the ETD1 method using the Léja point method and Krylov subpace technique (with $m = 6$).

We take the anisotropic permeability tensor $\mathbf{k}$ to be

$$\mathbf{k} = \begin{pmatrix} k_1 & 0 & 0 \\ 0 & k_2 & 0 \\ 0 & 0 & k_3 \end{pmatrix} \tag{3.86}$$

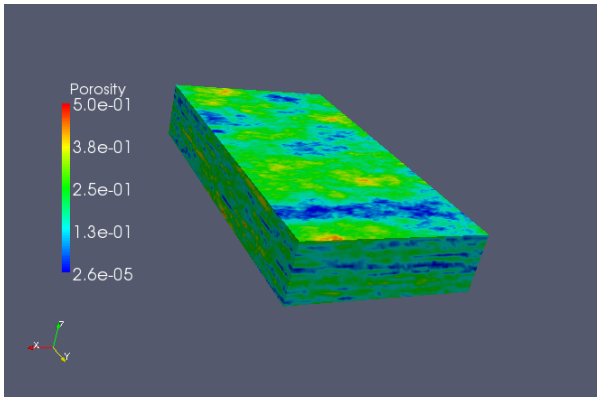
with $k_1 > 0, k_2 > 0, k_3 > 0$. We used the EEM scheme for the reference solution for ETD1 and semi implicit schemes. For the EEM scheme, the reference solution ("the true solution") is the numerical solution with the smallest Δt.

To analyse the efficiency of the ETD1 and EEM schemes compared to standard semi-implicit time integrators, we use the upper 20 layers (Example 1) and upper 40 layers (Example 2) of the highly heterogeneous SPE10 case (Figure 3.8), which represents a braided fluvial North Sea oil field with seven orders of magnitude permeability variation [1]. We consider two cases for dispersion. First, we take a uniform dispersion tensor of $\mathbf{D} = 10^{-6} \times \mathbf{I}_3$ and secondly we take $\mathbf{D} = (D_{i,j})$ as function of the velocity $\mathbf{q}$

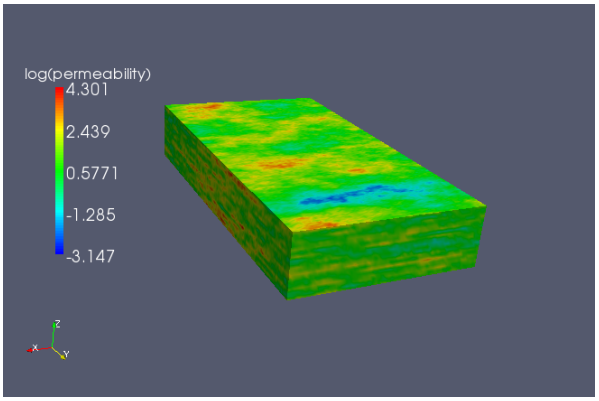
$$D_{i,j} = \alpha_T \|\mathbf{q}\| \delta_{i,j} + (\alpha_L - \alpha_T) q_i q_j / \|\mathbf{q}\|, \quad 1 \le i, j \le 3, \qquad \alpha_T = \alpha_L = 1, \tag{3.87}$$

where α_T and α_L are the longitudinal and the transverse dispersivity, respectively.

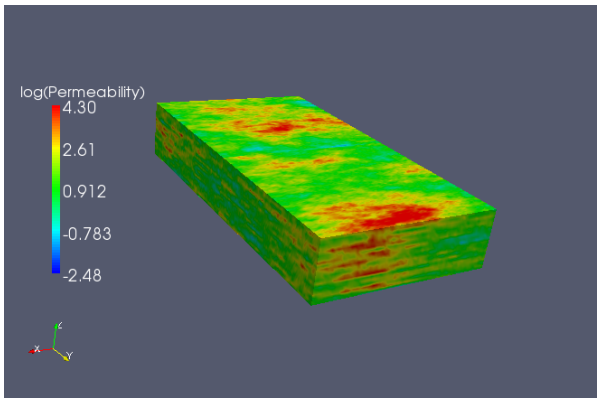
All our tests were performed on a workstation with a 3 GHz Intel processor and 8GB RAM. Our code was implemented in Matlab 7.10. In contrast to our earlier 2D simulations,



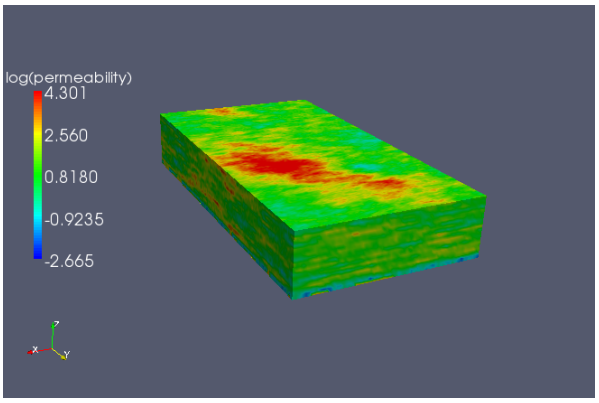
(a)



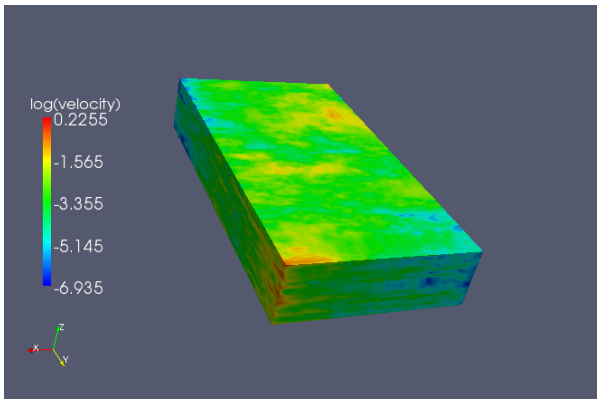
(b)



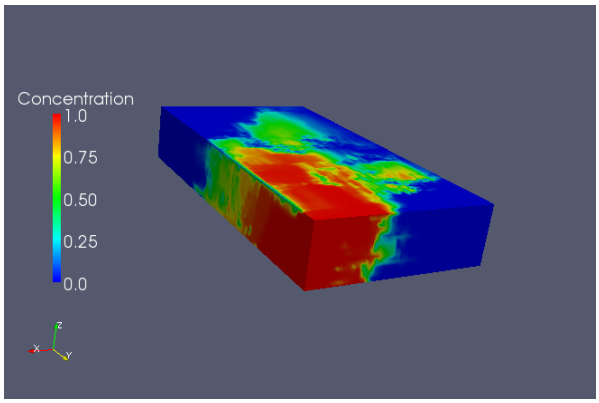
(c)



(d)



(e)



(f)

Figure 3.8: Porosity (a) and permeability in $x-$, $y-$, and $z-$direction (b-d, respectively), norm of the velocity field (e), and concentration after $T = 25600$ seconds (f) for the first upper 20 layers of the SPE 10 model [1]. Note that the vertical depth is exaggerated ten-fold. The injector is located in the lower left corner (high concentration) and the producer diagonally opposite in the upper right corner (low concentration). The ADR (2.8) was solved without chemical reactions, constant dispersion tensor, and a tolerance of $\varepsilon = 10^{-6}$. Note that the results of the EEM and ETD1 scheme are identical for the same tolerance as both schemes solve the ADR exactly in time within the given tolerance.

we now use large time-steps, very large Péclet numbers, and an iterative solver for linear systems. Further, in contrast to the simulations presented in [64] for the EEM scheme used in conjunction with a 2D finite element method, we use the Krylov subspace technique and large Péclet number flows. The resulting domain is $\Omega = [0, L_1] \times [0, L_2] \times [0, L_3]$, the finite volume mesh $\mathcal{T}$ has a spatial discretization $\Delta x = 20$ ft, $\Delta y = 10$ ft, and $\Delta z = 2$ ft. Dimensions are $L_1 = 1200$ ft, $L_2 = 2200$ ft, and $L_3 = 20$ ft in Example 1 and $L_3 = 40$ ft in Example 2.

The matrix A_h is sparse with size $264,000 \times 264,000$ and $1,810,400$ non-zero elements for Example 1 and $528,000 \times 528,000$ with $3,647,200$ non-zero elements for Example 2. For the semi-implicit time integration, this linear system is solved at each time-step with a variant of the iterative Krylov solver, the Bi-Conjugate Gradients Stabilized Method (Bi-CGStab) as implemented in Matlab. We use $\varepsilon = 10^{-6}$ as the absolute tolerance error and $m = 8$ for the Krylov subspace dimension. To our knowledge, there is no rigorous theory that allows us to predict the optimal value for m a priori. For small m, a penalty can arise from an increase in the number of iterations necessary to achieve a given tolerance, especially if Δt is large, but less time is spent in the orthogonalisation process and the required memory is lower. When m increases, the total number of iterations decreases but a penalty occurs due to the additional time spent in the orthogonalisation process and the corresponding increase in memory requirements.

For pressure and concentration, we take the Dirichlet boundary condition

$$\Gamma_D = \{\{0\} \times \{0\} \times [0, L_3]\} \cup \{\{L_1\} \times \{L_2\} \times [0, L_3]\},$$

and homogenous Neumann boundary conditions elsewhere such that

$$\begin{aligned} p &= \begin{cases} 3998.96 \ \text{psi} \quad \text{in} \quad \{0\} \times \{0\} \times [0, L_3] \\ \\ 7997.92 \ \text{psi} \quad \text{in} \quad \{L_1\} \times \{L_2\} \times [0, L_3] \end{cases} \\ -\mathbf{k}\,\nabla p(\mathbf{x}, t) \cdot \mathbf{n} &= 0 \quad \text{in} \quad \Gamma_N = \partial\Omega \backslash \Gamma_D. \end{aligned}$$

This models fixed-pressure injector and producer located at $\{0\} \times \{0\} \times [0, L_3]$ and $\{L_1\} \times \{L_2\} \times [0, L_3]$, respectively.

For the concentration we take

$$\begin{aligned} X &= 0 \quad \text{in} \quad \{\{0\} \times \{0\} \times [0, L_3]\} \times [0, T]\,, \\ X &= 1 \quad \text{in} \quad \{\{L_1\} \times \{L_2\} \times [0, L_3]\} \times [0, T]\,, \end{aligned}$$

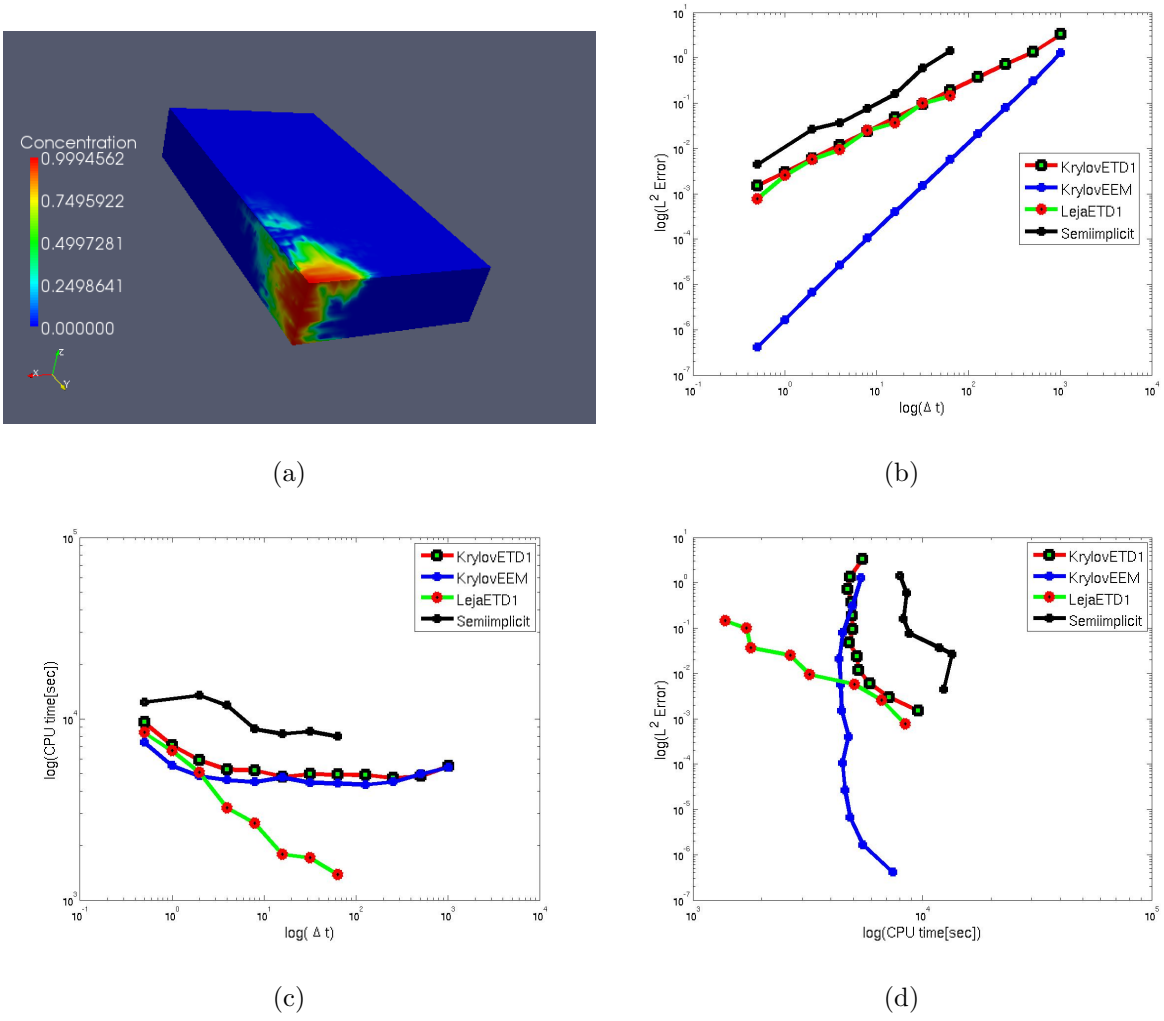
$$\begin{aligned} -(\mathbf{D}\nabla X)(\mathbf{x}, t) \cdot \mathbf{n} &= 0 \quad \text{in} \quad \Gamma_N \times [0, T]\,, \\ X_0 &= 0 \quad \text{in} \quad \Omega \quad \text{(initial solution)} \end{aligned}$$

where $\mathbf{n}$ is the unit outward normal vector to Γ_N. For the reaction function we use the classical Langmuir sorption isotherm given by $R(X) = (\lambda\beta X)/(1 + \lambda X)$, with $\lambda = 1$, $\beta = 10^{-3}$. The dynamic viscosity μ is $\mu = 0.3$ cp.

In the legends of the graphs shown below, "KrylovETD1" denotes results from the ETD1 scheme with the Krylov subspace technique, "LejaETD1" denotes results from the ETD1 scheme with the real fast Leja points technique, "Semiimplicit" denotes results from the standard semi-implict scheme, "KrylovEEM" denotes results from EEM scheme the Krylov subspace technique.

3.6.1 Example 1

Figure 3.9(a) shows the concentration field at time $T = 4096$ seconds for the first 20 upper layers of the SPE 10 model for the solution of the ADR with a constant diffusion tensor and with chemical reactions modelled by a Langmuir adsorption isotherm. As to be expected, flow follows regions of high porosity and permeabilities. Figure 3.9(b) shows the corresponding L^2 error as a function of the size of the time-step. It demonstrates that the ETD1 scheme is more accurate than the standard semi-implicit method. It further shows that the EEM scheme is also more accurate than the semi-implicit and ETD1 schemes. All three time integrators, semi-implicit, ETD1, and EEM, exhibit convergence that is in agreement with theory: the temporal order of the semi-implicit and ETD1 scheme is $\mathcal{O}(\Delta t)$ while the temporal order of the EEM scheme is $\mathcal{O}(\Delta t^2)$, hence the smallest error for a given time-step size is obtained with the EEM scheme. Figure 3.9(c) and Figure 3.9(d) demonstrate the



(a) (b)

(c) (d)

Figure 3.9: Concentration after $T = 4096$ seconds (a) for the first upper 20 layers of the SPE 10 model [1], L^2 error as a function of size of the time-step (b), CPU time as a function of size of the time-step (c), CPU time as a function of the L^2 error (d). Note that the vertical depth is exaggerated tenfold. The maximum local Péclet number is 1.7×10^6. The ADR (2.8) was solved with a constant dispersion tensor and with chemical reactions represented by a Langmuir adsorption isotherm.

efficiency of ETD1 and EEM schemes compared to the standard semi-implicit method by plotting the CPU time as a function of the size of the time-step, respectively the L^2 error as a function of the CPU time. They further show that the real Léja points technique is computationally more efficient than the Krylov subspace technique in the ETD1 scheme and a slight improvement in efficiency can be obtained with the EEM scheme compared to the ETD1 scheme with the Krylov subspace technique. Note there is a plateau in the CPU time for the Krylov subspace methods (Figure 3.9(c) and Figure 3.9(d)). This can be explained by the fact that each time-step is subdivided into smaller sub-steps to reach a given tolerance ε in the function `phiv.m` of the package Expokit [50], providing a limit for the efficiency of this method (see also Figure 3.10(c) and Figure 3.10(d)). Another important observation is that for the ETD1 scheme with the real Léja technique, the CPU time appears to be proportional to the size of the time-step, indicating that the best efficiency is reached for large time-steps sizes. Generally, these results imply that the best efficiency of the ETD1 and EEM schemes is obtained for the largest time-steps, although at the cost of accuracy.

Figure 3.10(a) shows the concentration field at time $T = 256$ seconds for the first 20 upper layers of the SPE 10 model for the solution of the ADR with a heterogeneous dispersion tensor (3.87) and with chemical reactions modelled by a Langmuir adsorption isotherm.

Plots of L^2 error and CPU time versus the size of the time-step (Figure 3.10(b) and Figure 3.10(c)) and the cross-plot of L^2 error versus CPU time (Figure 3.10(d)) show that, like in Figure 3.9, the EEM scheme is more accurate than the ETD1 and semi-implicit schemes as it is of order $\mathcal{O}(\Delta t^2)$. However, for this case the EEM scheme is significantly less efficient than the ETD1 scheme with the Léja points technique. Likewise, the ETD1 scheme with the Krylov subspace technique is less efficient than the Léja points technique and both Krylov subspace methods show flat lines in plots of CPU time versus time-step size and L^2 error versus CPU time. It hence appears that a heterogeneous and anisotropic dispersion tensor reduces the efficiency of the ETD1 and EEM schemes and smaller time-steps may help to increase the efficiency; however, both schemes are still more accurate than the semi-implicit time integrator. As explained above, this is due to the time-step being subdivided into smaller sub-steps to reach a given tolerance ε in the function `phiv.m`

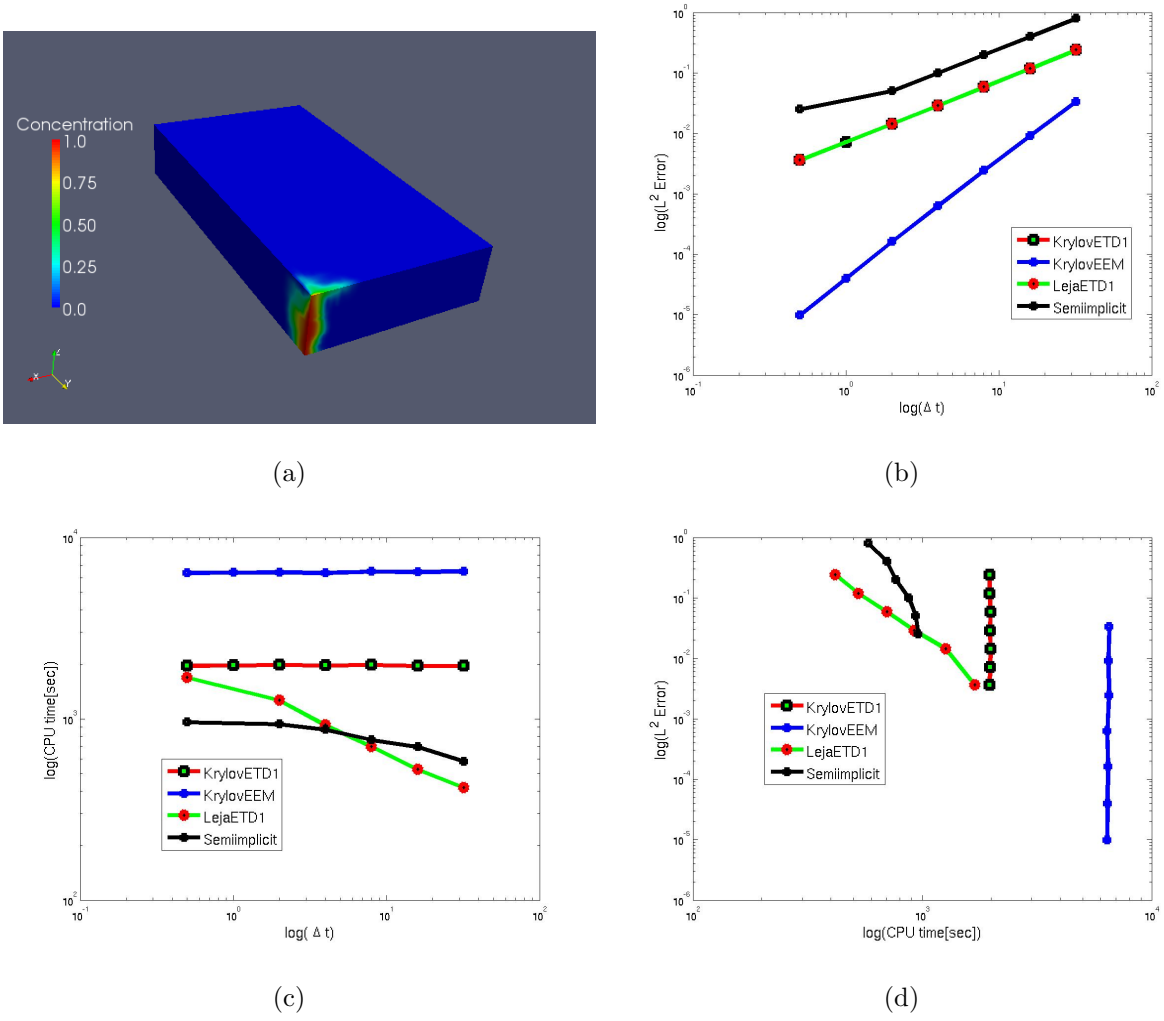


Figure 3.10: Concentration after $T = 256$ seconds (a) for the first upper 20 layers of the SPE 10 model [1], L^2 error as a function of size of the time-step (b), CPU time as a function of size of the time-step (c), CPU time as a function of the L^2 error (d). Note that the vertical depth is exaggerated tenfold. The maximum Péclet number is 2.4×10^6. The ADR (2.8) was solved with a heterogeneous dispersion tensor (3.87) and with chemical reactions represented by a Langmuir adsorption isotherm.

of the package Expokit [50]. The ETD1 method with the Léja points technique still shows scalability, i.e. the CPU time decreases with decreasing size of the time-step.

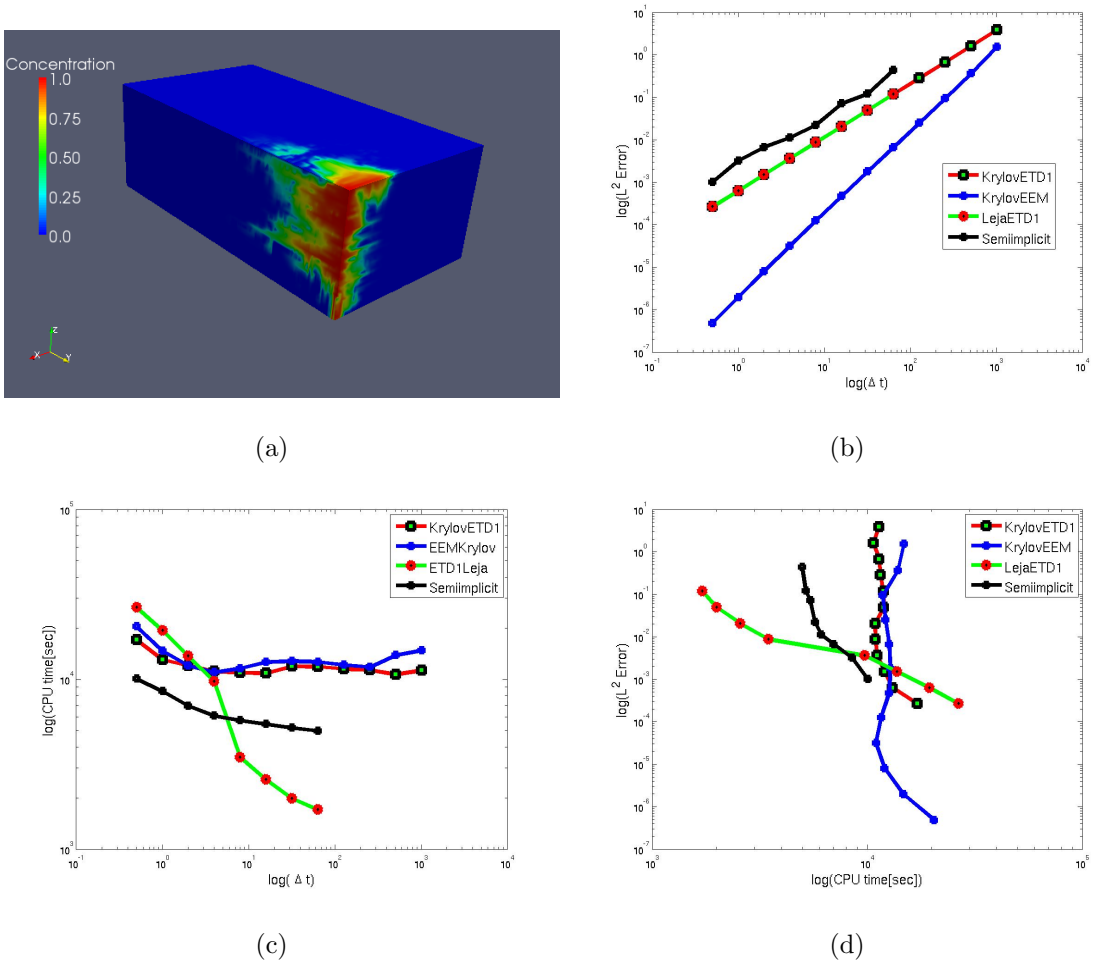


Figure 3.11: Concentration after $T = 4096$ seconds (a) for the first upper 40 layers of the SPE 10 model [1], L^2 error as a function of size of the time-step (b), CPU time as a function of size of the time-step (c), CPU time as a function of the L^2 error (d) . Note that the vertical depth is exaggerated tenfold. The maximum Péclet number is 3.2×10^6. The ADR (2.8) was solved with a homogeneous dispersion tensor and with chemical reactions represented by a Langmuir adsorption isotherm.

3.6.2 Example 2

Figure 3.11(a) shows the concentration field at time $T = 4096$ seconds for the first 40 upper layers of the SPE 10 model for the solution of the ADR with a homogeneous dispersion

tensor and chemical reactions modelled by a Langmuir adsorption isotherm. As before, ETD1 and semi-implicit schemes are of order $\mathcal{O}(\Delta t)$ while the EEM scheme is of $\mathcal{O}(\Delta t^2)$ (Figure 3.11(b)). Figure 3.11(c) shows that the ETD1 scheme with the Léja points technique is only more efficient than the standard semi-implicit scheme for large time-steps. For smaller time-steps, the ETD1 scheme with the Léja points technique becomes less efficient because more Léja points are needed for the interpolation. This is probably due to the lack of accuracy in the computation of the divided differences (3.73). We can recover the efficiency by adding more terms in the Taylor expansion during the computation of the divided differences (3.73) or by scaling the matrix $\mathbf{L}_m$ to smaller entries when used in the algorithm of [64] to compute the finite differences. Scaling the matrix $\mathbf{L}_m$ renders the "squaring" procedure less efficient, but as the divided differences are computed only once per time-step, this will not increase the overall computational cost greatly. Figure 3.11(c) further shows that the efficiency of the ETD1 and EEM schemes with the Krylov subspace technique is similar. Like in the previous two examples, Figure 3.11(d) shows that the subdivision of the time-step into smaller sub-steps in the function `phiv.m` of the package Expokit [50] reduces the efficiency of the Krylov subspace based techniques EEM and ETD1 for large time-steps.

3.7 Concluding remarks

We have developed two exponential time integrators of order one (ETD1) and order 2 (EEM) where the matrix exponential function φ_1 is computed with either real fast Léja points techniques or the Krylov subspace technique. In 2D, we have applied the scheme ETD1 to a variety of linear and non-linear advection-diffusion-reaction problems in homogeneous as well as highly heterogeneous porous media where the spatial discretisation was achieved by standard upwind-weighted finite volume meshes on non-uniform rectangular grids. The largest problems comprised 10^6 unknowns. We compared the performance of the ETD1 method to standard semi-implicit and implicit time integrators. Transport in our example applications was advection as well as diffusion dominated. All our numerical examples demonstrate that the ETD1 scheme is highly competitive compared to standard time integrators. This competitiveness comprises two components: efficiency and accuracy.

Generally, the ETD1 method requires at least 10 times less computational cost compared to implicit time integrators to reduce the numerical error to a certain value. Semi-implicit time integrators perform at best similar to our ETD1 method. The real fast Léja points technique is on average equivalent to or more efficient than the Krylov subspace technique. A similar observation was made in Martinez et al. [55] and Bergamaschi et al. [56] for example applications with constant dispersion tensors, uniform velocity fields, and low Péclet number flows, where the spatial discretisation was achieved by finite difference and finite element space discretisation. A single computation of ETD1 with real fast Léja points requires a few seconds on a standard PC, even with our uncompiled Matlab code.

In 3D, we have implemented ETD1 and EEM schemes where we computed the matrix exponential function φ_1 either with the real fast Léja points techniques (ETD1 scheme) or with the Krylov subspace technique (ETD1 and EEM schemes). We have applied them to a variety non-linear advection-diffusion-reaction problems in highly heterogeneous 3D porous media based on the SPE 10 test case [1]. The spatial discretisation was achieved by standard upwind-weighted finite volume method and the resulting matrices contained several 100k unknowns. Fluid flow is dominated by advection (Péclet numbers larger than 10^6). Simulations were carried out on a standard workstation with a 3 GHz Intel processor and 8 GB RAM. Our code was implemented in Matlab 7.10. Our analysis showed that both exponential time integrator schemes outperformed the standard semi-implicit scheme in terms of efficiency and accuracy. As for 2D applications, we observed that the Léja points technique is on average equivalent to or more efficient than the Krylov subspace technique and that the ETD1 scheme with the Léja points technique scales well with the size of the time-step. The reduced performance of the ETD1 and EEM schemes with the Krylov subspace technique is caused by the function `phi.m` of the Expokit package [50], which creates internal time subdivisions to satisfy a given tolerance. If a large time-step is chosen, this subdivision can dominate the calculation and increase the CPU time, rendering the Krylov-based techniques less efficient. In general, our results suggest that the exponential integrators can readily be applied to large-scale 3D reservoir simulations with several million of unknowns as they always outperform semi-implicit time integrators. It hence may become a viable alternative to other scalable solvers such as hierarchical algebraic multigrid methods [69] or multi-scale methods (e.g, [70–72]) which are commonly used for

large-scale simulations of flow and transport in heterogeneous porous media.

Chapter 4

Background to nonlinear SPDEs and time discretizations

In this chapter, we give a rigorous introduction to nonlinear SPDEs and some numerical stochastic schemes. More details can be found in [26, 73, 74].

4.1 Existence and uniqueness

There are basically three approaches to analyse SPDEs, the martingale approach [26], the semigroup (or mild solutions approach) [26] and the variational approach [73]. Under some technical conditions (see [73]) solutions in these approaches are identical. In this thesis, we study the mild solutions to build new numerical schemes.

Let us give some basic definitions. In all this chapter H is a separable Hilbert space with norm $\|.\|$ and $(\mathbb{D}, \mathbb{A}, P)$ is a probability space. $\mathbb{D}$ is the sample space, $\mathbb{A}$ a $\sigma-$algebra and P a probability measure.

4.1.1 Basic definitions

Definition 4.1 ***[Measurability [26]]***
A function $X : \mathbb{D} \to H$ is $\mathbb{A}-$measurable if

$$X^{-1}(O) := \{\omega \in \mathcal{D},\ X(\omega) \in O\} \in \mathbb{A} \quad \textit{for all Borel sets } O \subset H. \tag{4.1}$$

Definition 4.2 ***[Filtration [26]]***

A filtration is an increasing sequence of $\sigma-$algebras $(\mathcal{F}_t)$, $t \in [0,T]$ of $\mathbb{A}$. The space $(\mathbb{D}, \mathbb{A}, (\mathcal{F}_t), P)$ is called a filtered probability space.

A filtration $(\mathcal{F}_t)$, $t \in [0,T]$ is normal if $\mathcal{F}_0$ contains all elements $O \in \mathbb{A}$ with $P(O) = 0$, and

$$\mathcal{F}_t = \underset{s>t}{\cap} \mathcal{F}_s \text{ for all } 0 \leq t \leq T.$$

Definition 4.3 ***[Adapted process [26]]***

A $H-$valued process $(Y(t))_{t\geq 0}$ is an adapted stochastic process to the filtration $(\mathcal{F}_t)_{t\geq 0}$ if $Y(t)$ is $\mathcal{F}_t$-measurable for all $t \geq 0$.

Let U be a separable space, then we denote by $L(U,H)$ the set of all linear bounded operators from U to H.

Definition 4.4 ***[Elementary process [26]]***

An $L(U,H)$-valued process $\Phi(t)$, $t \in [0,T]$ on $(\mathbb{D}, \mathbb{A}, P)$ with normal filtration $(\mathcal{F}_t)$ is elementary if there exists $0 = t_0 < \cdots < t_k = T$, $k \in \mathbb{N}$ such that

$$\Phi(t) = \sum_{m=0}^{k-1} \Phi_m 1_{[t_m, t_{m+1}]}(t), \ t \in [0,T],$$

where

- *$\Phi_m : \mathbb{D} \to L(U,H)$ is $\mathcal{F}_{t_m}-$measurable, with the strong Borel $\sigma-$algebra ($\sigma-$algebra generated by the open sets) on $L(U,H)$, $0 \leq m \leq k-1$,*
- *Φ_m takes only a finite number of values in $L(U,H)$, $0 \leq m \leq k-1$.*

We denote by $\mathcal{E}$ the class of elementary $L(U,H)-$valued processes.

Ito and Stratonovich integrals are related (see [74]). In this thesis we will use only the Ito integral.

Let Q be a positive, symmetric, bounded linear operator in H.

Definition 4.5 ***[$Q-$Wiener process [26]]***

A $H-$valued stochastic process $W(t)$, $t \in [0,T]$, on a probability space $(\mathbb{D}, \mathbb{A}, P)$ is called a $Q-$Wiener process if

- *$W(0) = 0$, $P-$a.s.*

- *W has $P-$a.s. continuous trajectories,*
- *the increments of W are independent, i.e. the random variables*

$$W(t_1), W(t_2) - W(t_1), \cdots, W(t_n) - W(t_{n-1})$$

are independent for all $0 \leq t_1 < \cdots < t_n \leq T,\ n \in \mathbb{N}$,

- *the increments have the following Gauss law*

$$P \circ (W(t) - W(s))^{-1} = N(0, (t-s)Q) \quad \textit{for all } 0 \leq s \leq t \leq T. \tag{4.2}$$

More information can be found in [26, 73].

Proposition 4.6 ***[Representation of the $Q-$Wiener process [73]]***
Let $e_k,\ k \in \mathbb{N}$, be an orthonormal basis of H consisting of eigenvectors of Q with corresponding eigenvalues $q_k,\ k \in \mathbb{N}$. Then a $H-$valued stochastic process $W(t),\ t \in [0,T]$, is a $Q-$Wiener process if and only if

$$W(t) = \sum_{i \in \mathcal{I}} \sqrt{q_i} e_i \beta_i(t), \tag{4.3}$$

where β_i are independent and identically distributed standard Brownian motions on a probability space $(\mathbb{D}, \mathbb{A}, P)$.

The linear and bounded operator Q is also called the covariance operator of the Wiener process W. If $Q = I$, then $\mathrm{Tr}(Q) = +\infty$ and the Wiener process is called cylindrical Wiener process and if $\mathrm{Tr}(Q) < \infty$, the process is called nuclear Wiener process.

Definition 4.7 ***[Ito and Stratonovich integrals [26, 74]]***

- *For an elementary process $\Phi(t)$, the Ito integral is defined as*

$$\int_0^t \Phi(s) dW(s) := \sum_{m=0}^{k-1} \Phi_m \left(W(t_{m+1} \wedge t) - W(t_m \wedge t) \right) \quad t \in [0,T], \tag{4.4}$$

and the Stratonovich integral as

$$\int_0^t \Phi(s) \circ dW(s) := \sum_{m=0}^{k-1} \frac{1}{2} \left(\Phi_{m+1} + \Phi_m \right) \left(W(t_{m+1} \wedge t) - W(t_m \wedge t) \right) \quad t \in [0,T], \tag{4.5}$$

where $t_m \wedge t = \min(t_m, t)$.

- *The definition of the Ito and Stratonovich integrals are extended to the completion* $\overline{\mathcal{E}}$ *of* $\mathcal{E}$ *in the space of* $L(U,H)-$*valued processes as the limits of Ito and Stratonovich integrals of a sequence of elementary processes i.e. for* $\Phi \in \overline{\mathcal{E}}$

$$\int_0^t \Phi(s)dW(s) := \lim_{n\to\infty} \int_0^t \Phi_n(s)dW(s), \tag{4.6}$$

$$\int_0^t \Phi(s) \circ dW(s) := \lim_{n\to\infty} \int_0^t \Phi_n(s) \circ dW(s) \tag{4.7}$$

where (Φ_n) *is a sequence of elementary processes with* $\Phi_n \to \Phi$ *as* $n \to \infty$*. The space* $\overline{\mathcal{E}}$ *belongs to the large set of predictable processes [73]. The space* $\overline{\mathcal{E}}$ *is also called the space of integrable and predictable processes.*

Definition 4.8 ***[Hilbert-Schmidt operator]***

An operator $T \in L(H) := L(H,H)$ *is Hilbert-Schmidt if*

$$\|T\|_{HS}^2 := \sum_{i\in\mathcal{I}} \|Te_i\|^2 < \infty,$$

where we denote by $\|.\|_{HS}$ *the Hilbert-Schmidt norm. The sum in* $\|.\|_{HS}^2$ *is independent of the choice of the orthonormal basis in* H*.*

We denote the space of Hilbert-Schmidt operators from $Q^{1/2}(H)$ to H by $L_2^0 := HS(Q^{1/2}(H), H)$ and the corresponding norm $\|.\|_{L_2^0}$ by

$$\|\varphi\|_{L_2^0} := \|\varphi Q^{1/2}\|_{HS} = \left(\sum_{i\in\mathcal{I}} \|\varphi Q^{1/2} e_i\|^2\right)^{1/2}.$$

Theorem 4.9 ***[Ito isometry [26]]***

Let φ *be a continuous* L_2^0-*process. We have the following equality known as the Ito's isometry*

$$\boldsymbol{E}\| \int_0^t \varphi dW\|^2 = \int_0^t \boldsymbol{E}\|\varphi\|_{L_2^0}^2 ds = \int_0^t \boldsymbol{E}\|\varphi Q^{1/2}\|_{HS}^2 ds.$$

Definition 4.10 ***[Ornstein-Uhlenbeck process]***

Let $k > 0$, $\sigma \geq 0$*. A* $\mathbb{R}-$*process* X *is called an Ornstein-Uhlenbeck process with mean* 0*, mean reversion rate* k *and volatility* σ*, if it satisfies the stochastic differential equation*

$$dX(t) = -kX(t)dt + \sigma d\beta(t) \tag{4.8}$$

where $\beta(t)$ *is the standard Brownian motion.*

This is a Gaussian process with the mild solution

$$X(t) = e^{-kt}X(0) + \sigma \int_0^t e^{k(s-t)} d\beta(s). \tag{4.9}$$

Applying the Ito isometry yields the following variance of $X(t)$

$$\mathrm{Var}(X(t)) = \frac{\sigma^2}{2k}\left(1 - e^{-2\,k\,t}\right). \tag{4.10}$$

4.1.2 Mild solution of semi linear SPDEs

The general Ito stochastic partial differential equation is given by

$$dX = (AX + F(X))dt + B(X)dW, \qquad X(0) = X_0 \in H, \qquad t \in [0,T], \;\; T > 0 \tag{4.11}$$

where A is a linear operator, generally unbounded, acting on a Hilbert space H, $F : H \to H$ and $B : H \to L_2^0$ are in general nonlinear, non continuous. The existence and uniqueness of the mild solution is ensured in [26,73,74] using the fixed point method under the assumption that the operator A is the generator of an analytic semigroup $S(t) = e^{tA}, t \geq 0$.

Definition 4.11 ***[Mild solution [26]]***

A predictable $H-$valued process $X(t)$, is said to be a mild solution of (4.11) if

$$P\left(\int_0^t \|X(s)\|^2 ds < +\infty\right) = 1 \tag{4.12}$$

and, for arbitrary $t \in [0,T]$ we have

$$X(t) = S(t)X_0 + \int_0^t S(t-s)F(X(s))ds + \int_0^t S(t-s)B(X(s))dW(s). \tag{4.13}$$

Our interest in this thesis is the parabolic SPDEs with nuclear covariance operator Q. For existence and uniqueness, we need the following assumptions.

Assumption 4.12 ***[Assumption on the drift term F]***

There exists a positive constant $L > 0$ such that F is continuous in H and satisfies the following condition

$$\|F(Z) - F(Y)\| \leq L\|Z - Y\| \qquad \forall Z, Y \in H,$$

As a consequence, there exists a constant $C > 0$ such that

$$\|F(Z)\| \leq \|F(0)\| + \|F(Z) - F(0)\| \leq \|F(0)\| + L\|Z\| \leq C(1 + \|Z\|).$$

Assumption 4.13 ***[Assumption on the diffusion term*** B ***for multiplicative noise]***
The covariance operator Q is nuclear and there exists a positive constant $L > 0$ such that B is continuous in H and satisfies the following condition

$$\|B(Z) - B(Y)\|_{L_2^0} \leq L\|Z - Y\| \qquad \forall Z, Y \in H.$$

As a consequence, the exists a constant $C > 0$ such that

$$\|B(Z)\|_{L_2^0} \leq \|B(0)\|_{L_2^0} + \|B(Z) - B(0)\|_{L_2^0} \leq \|B(0)\|_{L_2^0} + L\|Z\| \leq C(1 + \|Z\|).$$

Assumption 4.14 ***[Assumption on the noise for additive noise]***
The covariance operator Q is nuclear i.e. the noise is trace class, thus

$$Tr(Q) = \sum_{i \in \mathcal{I}} q_i < \infty.$$

Assumption 4.15 ***[Assumption on the linear operator*** A***]***
The operator A is the generator of an analytic semigroup $S(t) := e^{tA}, t \geq 0$.

Theorem 4.16 ***[Existence, uniqueness and properties of the mild solution [26]]***
Assume that the initial solution X_0 is an $\mathcal{F}_0-$measurable $H-$valued random variable and Assumption 4.12, Assumption 4.13 (or Assumption 4.14), Assumption 4.15 are satisfied.

- *There exists a mild solution X to (4.11) unique, up to equivalence among the processes satisfying*

$$P\left(\int_0^T \|X(s)\|^2 ds < \infty\right). \tag{4.14}$$

Moreover it has a continuous modification.

- *For any $p \geq 2$ there exists a constant $C = C(p,T) > 0$ such that*

$$\sup_{t\in[0,T]} \mathbf{E}\|X(t)\|^p \leq C\left(1 + \mathbf{E}\|X_0\|^p\right). \tag{4.15}$$

- *For any $p > 2$ there exists a constant $C_1 = C_1(p,T) > 0$ such that*

$$\mathbf{E}\sup_{t\in[0,T]} \|X(t)\|^p \leq C_1\left(1 + \mathbf{E}\|X_0\|^p\right). \tag{4.16}$$

4.2 Numerical schemes for SPDEs

The study of numerical solutions of SPDEs is an active area of research and there is a growing literature on numerical methods for SPDEs. For temporal discretizations, the linear implicit Euler scheme is often used [75,76], spatial discretizations are usually achieved with finite element [77–79], finite difference [75, 76, 80] and spectral method [81, 82]. The finite element, finite difference or finite volume space discretizations are more useful for complex domains, general unbounded operators A, and yield the discrete form may be written as

$$dX^h = (A_h X^h + F_h(X^h))dt + B_h(X^h)dW^h, \tag{4.17}$$

where A_h is a non-diagonal operator, F_h, B_h and W^h are respectively the spatial projection of F, B and W. The standard time discretizations are the explicit Euler-Maruyama scheme

$$X^h_{m+1} = (I + \Delta t A_h)\, X^h_m + \Delta t F_h(X^h_m) + B_h(X^h_m)\left(W^h_{m+1} - W^h_m\right), \tag{4.18}$$

the semi implicit Euler-Maruyama

$$X^h_{m+1} = (I - \Delta t A_h)^{-1}\left[X^h_m + \Delta t F_h(X^h_m) + B_h(X^h_m)\left(W^h_{m+1} - W^h_m\right)\right], \tag{4.19}$$

and the Crank–Nicholson scheme

$$X^h_{m+1} = \left(I - \frac{\Delta t}{2}A_h\right)^{-1}\left[\left(I + \frac{\Delta t}{2}A_h\right)X^h_m + \Delta t F_h(X^h_m) + B_h(X^h_m)\Delta W^h_m\right], \tag{4.20}$$

with

$$\Delta W^h_m \;=\; W^h_{m+1} - W^h_m = \sqrt{\Delta t}\sum_{i\in\mathcal{I}} \sqrt{q_i} R_{i,m} e_i, \tag{4.21}$$

where $R_{i,m}$ are independent, standard normally distributed random variables with means 0 and variance 1.

If the operator A is self adjoint and positive definite, the Galerkin spectral discretization yields the diagonal discrete form

$$dX^N = (A_N X^N + F_N(X^N))dt + B_N(X^N)dW^N,$$

with $A_N = P_N A,\ F_N = P_N F,\ B_N = P_N B,\ W^N = P_N W$, P_N is the spectral projection defined for $u \in H$ by

$$P_N u = \sum_{i \in \mathcal{I}_N} (e_i, u) e_i, \tag{4.22}$$

(e_i) being the eigenfunctions of the operator A, $\mathcal{I}_N \subset \mathcal{I}$ is the set containing the first N elements of $\mathcal{I}$ and $(.)$ is the inner product of H. The exponential integrator scheme in [81] is given by

$$X^N_{m+1} = e^{\Delta t A_N} \left(X^N_m + \Delta t F(X^N_m) + B(X^N_m) \left(W^N_{m+1} - W^N_m \right) \right). \tag{4.23}$$

If the operator A is a self adjoint and positive definite, the exponential integrator proposed in [82] has improved convergence properties for additive noise. The scheme in [82] for $B = I$ is given by

$$X^N_{m+1} = e^{\Delta t A_N} X^N_m + \Delta t \varphi_1(\Delta t A_N) F_N(X^N_m) + \int_{t_m}^{t_{m+1}} e^{(t_{m+1}-s)A_N} dW^N, \tag{4.24}$$

where the process

$$\widehat{O}_m = \int_{t_m}^{t_{m+1}} e^{(t_{m+1}-s)A_N} dW^N \tag{4.25}$$

has the exact variance in each Fourier mode as an Ornstein–Uhlenbeck process (see (6.17)).

Chapter 5

A modified semi–implict Euler-Maruyama Scheme for finite element discretization of SPDEs

We consider the numerical approximation of a general second order semi–linear parabolic stochastic partial differential equation (SPDE) driven by additive space-time noise. We introduce a new scheme using a linear functional of the noise with a semi–implicit Euler–Maruyama method in time and a finite element method in space. Extension to finite differences or finite volumes on space would be possible. We consider noise that is white in time and either in H^1 or H^2 in space. We give the convergence proofs in the root mean square L^2 norm for a diffusion reaction equation and in root mean square H^1 norm in the presence of advection. We examine the regularity of the initial data, the regularity of the noise and errors from projecting the noise. We present numerical results for a linear reaction diffusion equation in two dimensions as well as a nonlinear example of a two-dimensional stochastic advection diffusion reaction equation. We see from both the analysis and numerics that we have better convergence properties than the standard semi–implicit Euler–Maruyama method. The results of this chapter are presented in our paper [18].

5.1 Introduction

In this chapter, we analyse the strong numerical approximation of the Ito stochastic partial differential equation defined in $\Omega \subset \mathbb{R}^d$. Boundary conditions on the domain Ω are typically Neumann, Dirichlet or some mixed conditions. We consider

$$dX = (AX + F(X))dt + dW, \qquad X(0) = X_0, \qquad t \in [0, T], \qquad T > 0 \tag{5.1}$$

in a Hilbert space $H = L^2(\Omega)$. Here we assume that A is the generator of an analytic semigroup $S(t) := e^{tA}, t \geq 0$ with eigenfunctions e_i and eigenvalues λ_i, $i \in \mathbb{N}^d$. F is a nonlinear function of X and possibly ∇X. The noise term $W(x,t)$ is a Q-Wiener process that is white in time and we assume some spatial regularity defined on a filtered probability space $(\mathbb{D}, \mathcal{F}, \mathbf{P}, \{F_t\}_{t\geq 0})$. The noise can be represented as a series in the eigenfunctions of the covariance operator Q and we assume for convenience that Q and A have the same eigenfunctions. The representation of the noise $W(x,t)$ is given in (4.3) with $\mathcal{I} = \mathbb{N}^d$.

Minimum assumptions on A, F and W are given in Chapter 4, and under these type of technical assumptions it is well known, (see Theorem 4.16 or [26, 73, 74] for results in a more general framework) that the unique mild solution is given by

$$X(t) = S(t)X_0 + \int_0^t S(t-s)F(X(s))ds + O(t) \tag{5.2}$$

with the stochastic process O given by the stochastic convolution

$$O(t) = \int_0^t S(t-s)dW(s). \tag{5.3}$$

Typical examples of the above type of equation are stochastic (advection) reaction diffusion equations where $A = \Delta$ arising from example in pattern formation in physics and mathematical biology. We illustrate our work with both a simple reaction diffusion equation where we can construct an exact solution

$$dX = (D\Delta X - \lambda X)\, dt + dW, \tag{5.4}$$

where λ is a constant, as well as the stochastic advection reaction diffusion equation

$$dX = \left(D\Delta X - \nabla \cdot (\mathbf{q}X) - \frac{X}{|X| + 1} \right) dt + dW \tag{5.5}$$

where $D > 0$ is the diffusion coefficient and $\mathbf{q}$ is the Darcy velocity field [20].

We analyse convergence of a finite element discretization in space combined with a semi-implicit discretization in time that uses a linear functional to approximate the noise. This approach extends the analysis of Jentzen [82–84] which is based on a Fourier spectral discretization. This discretization diagonalizes the linear operator A and then exploits the fact that in each Fourier mode the noise is an Ornstein–Uhlenbeck process and hence the variance is known. For complex domains or mixed boundary conditions a Fourier spectral approach is not feasible and, for example, finite element (finite difference or finite volume) discretization is preferred, however this destroys the diagonalization of the linear operator. Here, we perform our analysis for the case of finite elements and numerically we also look at finite volumes. Our work differs from other finite element discretizations [77, 85, 86] where the approach to the noise is to consider it directly in the finite element space. We follow more closely [76, 79, 87] and introduce a projection onto a finite number modes and a projection onto the finite element space. The aim is to gain the flexibility of the finite element or the finite volume discretization to deal with complex boundaries, mixed boundary conditions and upwinding to deal with advection.

We give convergence proofs in root mean square $L^2(\Omega)$ norm for reaction-diffusion equations and in root mean square $H^1(\Omega)$ norm for advection reaction-diffusion for spatially regular noise. The smoothing effect of the semigroup generated by the operator A in the SPDE (5.1) and various semigroup estimates play an important role in the proofs.

The chapter is organised as follows. In Section 5.2 we present the numerical scheme and assumptions that we make on the linear operator, nonlinearity and the noise. We then state and discuss our main results. In Section 5.3.2 and Section 5.3.3 we present the proof of the convergence theorems. We end by presenting some simulations in Section 5.4 these are applied both to a linear example where we can compute an exact solution as well as a more realistic model coming from model of the advection and diffusion of a solute in a porous media with a non-linear reaction term.

5.2 Numerical scheme and main results

Let us start by presenting briefly the notation for some function spaces. For a Banach space $\mathcal{V}$ we denote by $L^{(2)}(\mathcal{V})$ the set of bounded bilinear mapping from $\mathcal{V} \times \mathcal{V}$ to $\mathbb{C}$ and $L_2(\mathbb{D}, \mathcal{V})$

the space defined by

$$L_2(\mathbb{D},\mathcal{V}) = \left\{ v \text{ random variable with value in } \mathcal{V} : \ \mathbf{E}\|v\|_{\mathcal{V}}^2 = \int_{\mathbb{D}} \|v(\omega)\|_{\mathcal{V}}^2 d\mathbf{P}(\omega) < \infty \right\}. \quad (5.6)$$

Throughout the thesis we assume that Ω is bounded and has a smooth boundary or is a convex polygon. For convenience of presentation we take A to be a self adjoint second order operator as this simplifies the convergence proof. More precisely

$$A = \nabla \cdot \mathbf{D}\nabla(.) + D_{0,0}\mathbf{I} = \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i}\left(D_{i,j}\frac{\partial}{\partial x_j}\right) + D_{0,0}\mathbf{I}, \quad (5.7)$$

where we assume as in Chapter 2 that $D_{i,j} \in L^{\infty}(\Omega)$, and that there exists a positive constant $c_1 > 0$ such that (2.14) holds. According to Chapter 2, the linear operator A generate an analytic semigroup $S(t) = e^{tA}$. We introduce two spaces $\mathbb{H}$ and V where $\mathbb{H} \subset V$, that depend on the choice of the boundary conditions for the domain of the operator A and for test functions. For Dirichlet boundary conditions we let

$$V = \mathbb{H} = H_0^1(\Omega) = \{v \in H^1(\Omega) : v = 0 \ \text{ on } \ \partial\Omega\}.$$

For Robin boundary conditions (Neumann boundary condition being a particular case) $V = H^1(\Omega)$ and

$$\mathbb{H} = \left\{ v \in H^1(\Omega) : \partial v/\partial \nu_A + \sigma v = 0 \quad \text{on} \quad \partial\Omega \right\}.$$

For mixed boundary conditions $\mathbb{H}$ is defined as in (2.32) and $V = H^1(\Omega)$ (see [17] for details). Functions in $\mathbb{H}$ can satisfy the boundary conditions and with $\mathbb{H}$ in hand we can characterize the domain of the operator $(-A)^{r/2}$ and have the following norm equivalence [32, 88] for $r = 1, 2$

$$\|v\|_{H^r(\Omega)} \equiv \|(-A)^{r/2}v\| =: \|v\|_r \qquad \forall v \in \mathcal{D}((-A)^{r/2}) = \mathbb{H} \cap H^r(\Omega).$$

The existence and uniqueness of the mild solution of equation (5.1) given by (5.2) is ensured if Assumption 4.12 and Assumption 4.14 are satisfied.

We start by discretizing the SPDE (5.1) in time. As in (5.2), by splitting we have

$$X(t) = \overline{X}(t) + O(t), \quad (5.8)$$

then $\overline{X}$ is the solution of the random PDE part of (5.1) with

$$\overline{X}(t) = S(t)X_0 + \int_0^t S(t-s)F(X(s))ds.$$

This solution at time $t_m = m\Delta t,\ \Delta t > 0$ is given by

$$\overline{X}(t_m) = S(t_m)X_0 + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S(t_m - s)F(X(s))ds. \tag{5.9}$$

The semi–implicit approximation of $\overline{X}$ in time is given by

$$Z_m = (I - \Delta tA)^{-m} X_0 + \Delta t \sum_{k=0}^{m-1} (I - \Delta tA)^{-(m-k)} F(Z_k + O(t_k)). \tag{5.10}$$

where we generate $O(t_k)$ from (5.3) by

$$O(t_k) = e^{\Delta tA} O(t_{k-1}) + \int_{t_{k-1}}^{t_k} e^{(t_k - s)A} dW(s). \tag{5.11}$$

If we assume that Q has the same eigenfunctions as the linear operator A and we have diagonalized the operator A by a Galerkin-Fourier spectral method then (5.11) reduces to an Ornstein–Uhlenbeck process in each Fourier mode as in [82].

This can then be simulated numerically with the correct mean and variance in each mode. Here we examine how this improvement can be maintained using a finite–element discretization or other spatial discretization that does not diagonalize the operator A (such as finite volumes or finite differences).

We consider discretization of the spatial domain by a finite element triangulation. Let $\mathcal{T}_h$ be a set of disjoint intervals of Ω (for $d = 1$), a triangulation of Ω (for $d = 2$) or a set of tetrahedra (for $d = 3$). Let $V_h \subset V$ denote the space of continuous functions that are piecewise linear over the triangulation $\mathcal{T}_h$. To discretize in space we introduce two projections. Our first projection operator P_h is the $L^2(\Omega)$ projection onto V_h defined for $u \in L^2(\Omega)$ by

$$(P_h u, \chi) = (u, \chi) \qquad \forall\ \chi \in V_h. \tag{5.12}$$

Then $A_h : V_h \to V_h$ is the discrete analogue of A defined by

$$(A_h \varphi, \chi) = (A\varphi, \chi) \qquad \varphi, \chi \in V_h. \tag{5.13}$$

We denote by S_h the semigroup generated by the operator A_h

The second projection $P_N,\ N \in \mathbb{N}$ is the projection onto a finite number of spectral modes e_i defined in (4.22).

The semi–discrete in space version of the problem (5.1) is to find the process $X^h(t) = X^h(.,t) \in V_h$ such that for $t \in [0,T]$,

$$dX^h = (A_h X^h + P_h F(X^h))dt + P_h P_N dW, \qquad X^h(0) = P_h X_0. \tag{5.14}$$

We denote by $\overline{X}^h$ the solution of the random system

$$\overline{X}^h(t) = S_h(t)X^h(0) + \int_0^t S_h(t-s)F(X^h(s))ds.$$

We now discretize in time by a semi–implicit method to get the fully discrete approximation of $\overline{X}^h$ defined by Z_m^h

$$Z_m^h = S_{h,\Delta t}^m P_h X_0 + \Delta\, t \sum_{k=0}^{m-1} S_{h,\Delta t}^{(m-k)} P_h F(Z_k^h + P_h P_N O(t_k)). \tag{5.15}$$

where

$$S_{h,\Delta t} := (\mathrm{I} - \Delta t\ A_h)^{-1}. \tag{5.16}$$

It is straightforward to show that

$$Z_{m+1}^h = S_{h,\Delta t}\left(Z_m^h + \Delta t P_h F(Z_m^h + P_h P_N O(t_m))\right). \tag{5.17}$$

Finally we can define our approximation X_m^h to $X(t_m)$, the solution of equation (5.1)

$$X_m^h = Z_m^h + P_h P_N O(t_m). \tag{5.18}$$

Therefore

$$X_{m+1}^h = S_{h,\Delta t}\left(X_m^h + \Delta t\, P_h F(X_m^h) - P_h P_N O(t_m)\right) + P_h P_N O(t_{m+1}) \tag{5.19}$$

is the new numerical scheme with $O(t_{m+1})$ generated from (5.11). It uses a finite element discretization and projects a linear functional of the noise and hence we expect superior approximation properties.

Remark 5.1 *In the new numerical scheme in (5.19), one can substitute $S_{h,\Delta t}$ by the following rational approximation of the exponential operator using in the Crank–Nicholson scheme*

$$S_{h,\Delta t} = \left(I - \frac{\Delta t}{2}A_h\right)^{-1}\left(I + \frac{\Delta t}{2}A_h\right). \tag{5.20}$$

For convergence proofs below we need sufficient regularity of the mild solution X, and therefore we will use some weak assumptions.

We describe in detail the weak assumptions that we make on the linear operator A, on our finite element discretization, the nonlinear term F and the noise dW.

Assumption 5.2 *The linear operator $-A$ is positive definite. Then there exists sequences of positive real eigenvalues $\{\lambda_n\}_{n\in\mathbb{N}^d}$ with $\inf_{i\in\mathbb{N}^d}\lambda_i > 0$ and an orthonormal basis in H of eigenfunctions $\{e_i\}_{i\in\mathbb{N}^d}$ such that the linear operator $-A : \mathcal{D}(-A) \subset H \to H$ is represented as*

$$-Av = \sum_{i\in\mathbb{N}^d} \lambda_i (e_i, v) e_i \qquad \forall \quad v \in \mathcal{D}(-A)$$

where the domain of $-A$, $\mathcal{D}(-A) = \{v \in H : \sum_{i\in\mathbb{N}^d} \lambda_i^2 |(e_i, v)| < \infty\}$.

Assumption 5.3 ***[Nonlinearity]***

Let $\mathcal{V}$ be a separable Banach space such that $\mathcal{D}((-A)^{1/2}) \subset \mathcal{V} \subset H = L^2(\Omega)$ continuously. We assume that there exists a positive constant $L > 0$ such that F satisfies one of the following.

(a) $F : \mathcal{V} \to \mathcal{V}$ is a twice continuously Nemytskii Fréchet differentiable mapping with

$$\|F'(v)w\| \le L\|w\|, \qquad \|F'(v)\|_{L(\mathcal{V})} \le L, \qquad \|F''(v)\|_{L^{(2)}(\mathcal{V})} \le L$$

and

$$\|(F'(u))^*\|_{L(\mathcal{D}((-A)^{1/2}))} \le L(1 + \|u\|_{\mathcal{D}((-A)^{1/2})}) \qquad \forall\, v, w \in \mathcal{V}, \quad u \in \mathcal{D}((-A)^{1/2}),$$

where $(F'(u))^*$ is the adjoint of $F'(u)$ defined by

$$((F'(u))^* v, w) = (v, F'(u)w) \quad \forall\, v, w \in H = L^2(\Omega).$$

As a consequence

$$\|F(Z) - F(Y)\| \le L\|Z - Y\| \qquad \forall Z, Y \in H,$$

and $\forall\, Y \in H = L^2(\Omega)$

$$\|F(Y)\| \le \|F(0)\| + \|F(Y) - F(0)\| \le \|F(0)\| + L\|Z\| \le C(\|F(0)\| + \|Y\|).$$

(b) F is globally Lipschitz continuous from $(H^1(\Omega), \|.\|_{H^1(\Omega)})$ to $(H = L^2(\Omega), \|.\|)$ then

$$\|F(Z) - F(Y)\| \le L\|Z - Y\|_{H^1(\Omega)} \qquad \forall Z, Y \in H^1(\Omega).$$

Remark 5.4 *It is important to notice that when F satisfies Assumption 5.3 (b), the fixed point method needs to be applied in the Hilbert space $H^1(\Omega)$ for the existence and uniqueness of the solution of equation (5.1).*

We assume that the function F is defined in $L^2(\Omega)$, although in general F may be defined in any Hilbert space. The possible choice of $\mathcal{V}$ can be H, $H^1(\Omega)$ or $C(\Omega)$ if $d = 1$. Finally we note that condition (a) in Assumption 5.3 for the Nemytskii operator F has recently been used in [89].

We assume sufficient regularity of the noise for the existence of a mild solution and to project the noise into the space V_h. To be specific we assume that the stochastic process O is in H^1 or H^2 in space, useful for the finite element projection P_h in the errors estimates.

Notice that for all $t \in [0, T]$ the process $O(t)$ is an adapted stochastic process to the filtration $(\mathcal{F}_t)_{t\geq 0}$ with continuous sample paths such that $O(t_2) - S(t_2 - t_1)O(t_1),\ 0 \leq t_1 < t_2 \leq T$ is independent of $\mathcal{F}_{t_1}$.

Assumption 5.5 ***[Stochastic process** O**]*** *We assume that the stochastic process O satisfies one or both of the following.*

(a)

$$O(t) \in L_2(\mathbb{D}, \mathcal{D}((-A)^{r/2})),\ \ \mathcal{D}((-A)^{r/2}) = \mathbb{H} \cap H^r(\Omega), \qquad r = 1, 2.$$

(b) For some $\theta \in (0, 1/2]$ and positive constant $C > 0$

$$\mathbf{E}\left(\|O(t_2) - O(t_1)\|_{\mathcal{V}}^4\right) \ \leq \ C(t_2 - t_1)^{4\theta} \qquad 0 \leq t_1 < t_2 \leq T.$$

Assumption 5.5 implies the regularity of the noise W, **throughout this thesis we will make a slight abuse by saying " noise in H^r " when** $O(t) \in L_2(\mathbb{D}, H^r(\Omega)),\ \forall t \in [0, T]$. Using the equivalence of norms, we have that

$$O(t) \in L_2(\mathbb{D}, \mathcal{D}((-A)^{r/2})) \quad \forall t \in [0, T] \Leftrightarrow \|(-A)^{r/2} Q^{1/2}\|_{HS} < \infty \quad r = 1, 2.$$

One can prove that with this assumption, for $\mathcal{V} = H = L^2(\Omega)$ we can take $\theta = 1/2$. For $\mathcal{V} = H^1(\Omega)$ we can take $\theta = 1/2$ if $O(t) \in L_2(\mathbb{D}, \mathcal{D}(-A))$ and $\theta \neq 1/2$ but close to $1/2$ if $O(t) \in L_2(\mathbb{D}, \mathcal{D}((-A)^{1/2}))$. The proof is similar to that in Lemma 5.10.

Remark 5.6 *It is important to notice that if* $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$ *with* $\mathbf{E}\|(-A)^\gamma X_0)\|^4 < \infty$, *Assumption 5.2, Assumption 5.3 and Assumption 5.5 ensure the existence of the unique solution* $X(t) \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$ *such that*

$$\mathbf{E}\left(\sup_{0\leq s\leq T}\|(-A)^\gamma X(s))\|^4\right) < \infty.$$

In general if $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$ *with* $\mathbf{E}\|(-A)^\gamma X_0)\|^4 < \infty$, *and* $O(t) \in L_2(\mathbb{D}, \mathcal{D}((-A)^\alpha))$ *then*

$X \in L_2(\mathbb{D}, \mathcal{D}((-A)^{min(\gamma,\alpha)}))$ *with*

$$\mathbf{E}\left(\sup_{0\leq s\leq T}\|(-A)^{min(\gamma,\alpha)} X(s)\|^4\right) < \infty.$$

More information about properties of the solution of the SPDE (5.1) can be found in [89].

5.2.1 Main results

Throughout the chapter we let N be the number of terms of truncated noise, $\mathcal{I}_N = \{1, 2, ..., N\}^d$ and take $t_m = m\Delta t \in (0, T]$, where $T = M\Delta t$ for $m, M \in \mathbb{N}$. We take C to be a constant that may depend on T and other parameters but not on Δt, N or h. We also assume that when initial data $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$ then $\mathbf{E}\|(-A)^\gamma X_0\|^4 < \infty$, with $0 \leq \gamma < 1$.

Our first result is a strong convergence result in L^2 when the non-linearity satisfies the Lipschitz condition of Assumption 5.3 (a).

Theorem 5.7 *Suppose that Assumptions 5.2, 5.3(a) and 5.5 (with* $r = 1, 2$*) are satisfied and* $F(X(t)) \in L_2(\mathbb{D}, \mathbb{H}), \forall\, t \in [0, T]$ *with* $\sup_{0\leq s\leq T} \mathbf{E}\|F(X(s))\|_1^2 < \infty$. *Let* $X(t_m)$ *be the mild solution of equation (5.1) represented by (5.2) and* X_m^h *be the numerical approximation through (6.5). Let* $1/2 \leq \gamma < 1$ *and set* $\sigma = \min(2\theta, \gamma)$ *and let* $\theta \in (0, 1/2]$ *be defined as in Assumption 5.5. If* $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$ *then*

$$\left(\mathbf{E}\|X(t_m) - X_m^h\|^2\right)^{1/2} \leq C\left(t_m^{-1/2}(h^r + \Delta t^\sigma) + \Delta t\,|\ln(\Delta t)| + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-r/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ *then*

$$\left(\mathbf{E}\|X(t_m) - X_m^h\|^2\right)^{1/2} \leq C\left(h^r + \Delta t^{2\theta} + \Delta t\,|\ln(\Delta t)| + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-r/2}\right).$$

If we assume stronger regularity on the noise we can obtain a strong error estimate in the H^1 norm.

Theorem 5.8 *Suppose that Assumptions 5.2, 5.3(b), 5.5(a) (with $r = 2$) are satisfied and $F(X(t)) \in L_2(\mathbb{D}, \mathbb{H}), \forall\ t \in [0,T]$ with $\sup_{0\le s\le T} \mathbf{E}\|F(X(s)\|_1^2 < \infty$. Let X be the solution mild of equation (5.1) represented by equation (5.2). Then we have the following: If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ then*

$$(\mathbf{E}\|X(t_m) - X_m^h\|^2_{H^1(\Omega)})^{1/2} \ \le\ C\left((h + \Delta t^{1/2-\epsilon} t_m^{-1/2}) + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N} \lambda_j\right)^{-1/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ and $-AX_0 \in L_2(\mathbb{D}, H^1(\Omega))$ then

$$(\mathbf{E}\|X(t_m) - X_m^h\|^2_{H^1(\Omega)})^{1/2} \ \le\ C\left((h + \Delta t^{1/2-\epsilon}) + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N} \lambda_j\right)^{-1/2}\right),$$

for small enough $\epsilon \in (0, 1/2)$.

First note in both theorems we see that if the initial data are not sufficiently smooth then the error is dominated by this term. This behaviour is typical of non-smooth error estimates. Secondly we remark that if we denote by N_h the number of vertices in the finite element mesh then it is well known (see for example [79]) that if $N \ge N_h$ then

$$\left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N} \lambda_j\right)^{-1} \le Ch^2 \qquad \text{and} \qquad \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N} \lambda_j\right)^{-1/2} \le Ch.$$

As a consequence the estimates in Theorem 5.7 and Theorem 5.8 can be expressed as a function of h and Δt only and it is the error from the finite element approximation that dominates. If $N \le N_h$ then it is the error from the projection P_N of the noise onto a finite number of modes that dominates.

From Theorem 5.8 we also get an estimate in the root mean square $L^2(\Omega)$ norm in the case that the nonlinear function F satisfies Assumption 5.3 (b).

Finally we note that if the nodes of the finite element mesh coincide with evaluations of $O(x,t)$ then the projection operator P_h is trivial. This also leads to a computational advantage as we no longer need the projection, we comment further in Section 5.4.

5.3 Proofs of main results

5.3.1 Some preparatory results

We introduce the Riesz representation operator $R_h : V \to V_h$ defined by

$$(-AR_h v, \chi) = (-Av, \chi) = a(v, \chi) \qquad v \in V,\ \forall \chi \in V_h. \tag{5.21}$$

Under the regularity assumptions on the triangulation and in view of the $V-$ellipticity (2.28), it is well known (see [17]) that the following error bounds holds

$$\|R_h v - v\| + h\|R_h v - v\|_{H^1(\Omega)} \leq Ch^r \|v\|_{H^r(\Omega)}, \qquad v \in V \cap H^r(\Omega),\ \ r \in \{1, 2\}. \tag{5.22}$$

It follows that

$$\|P_h v - v\| \leq Ch^r \|v\|_{H^r(\Omega)} \qquad \forall v \in V \cap H^r(\Omega), \quad r = 1, 2. \tag{5.23}$$

We examine the deterministic linear problem. Find $u \in V$ such that

$$u' = Au \qquad \text{given} \quad u(0) = v, \quad t \in (0, T]. \tag{5.24}$$

The corresponding semi–discretization in space is : find $u_h \in V_h$ such that $u_h' = A_h u_h$ where $u_h^0 = P_h v$. The full discretization of (5.24) using implicit Euler in time is given by

$$u_h^{n+1} = (\mathrm{I} - \Delta t\ A_h)^{-(n+1)}\ P_h v = S_{h,\Delta t}^{n+1}\ P_h v.$$

We consider the error at $t_n = n\Delta t$ and define the operator T_n from

$$u(t_n) - u_h^n = (S(t_n) - (I - \Delta t\, A_h)^{-n}\ P_h)v := T_n v. \tag{5.25}$$

Lemma 5.9 *The following estimates hold on the numerical approximation to (5.24).*

Estimation in $H = L^2(\Omega)$ ***norm****. If* $v \in \mathbb{H}$ *then*

$$\|u(t_n) - u_h^n\| \quad = \quad \|T_n v\| \leq Ct_n^{-1/2}(h^2 + \Delta t)\|v\|_1 \tag{5.26}$$

and if $v \in \mathcal{D}(-A) = \mathbb{H} \cap H^2(\Omega)$

$$\|u(t_n) - u_h^n\| \quad = \quad \|T_n v\| \leq C(h^2 + \Delta t)\|v\|_2. \tag{5.27}$$

Estimation in $H^1(\Omega)$ ***norm****. If* $v \in \mathbb{H}$ *then*

$$\|u(t_n) - u_h^n\|_{H^1(\Omega)} \quad = \quad \|T_n v\|_{H^1(\Omega)} \leq C\|v\|_1(t_n^{-1/2}h + t_n^{-1}\Delta t). \tag{5.28}$$

If $v \in \mathcal{D}(-A) = \mathbb{H} \cap H^2(\Omega)$

$$\|u(t_n) - u_h^n\|_{H^1(\Omega)} \quad = \quad \|T_n v\|_{H^1(\Omega)} \leq C\|v\|_2(h + t_n^{-1/2}\Delta t). \tag{5.29}$$

Finally, if $v \in \mathcal{D}(-A)$ *and* $Av \in H^1(\Omega)$ *then*

$$\|u(t_n) - u_h^n\|_{H^1(\Omega)} \quad = \quad \|T_n v\|_{H^1(\Omega)} \leq C\| - Av\|_{H^1(\Omega)}(h + \Delta t). \tag{5.30}$$

Proof. We just give here some references for the proof. Estimates in the $H = L^2(\Omega)$ norm are given in [17, 31]. In [31], $A = \Delta$ with Dirichlet boundary conditions. Estimates in the $H^1(\Omega)$ norm are the special cases of Theorem 5.3 in [32] where the proof is given for a general semi–linear parabolic problem with a locally Lipschitz nonlinear term. To obtain our result from [32] note that $u(t) = S(t)v$ so that we have the analogue of [32, Theorem 5.2]

$$\|u_t(t)\|_{H^2(\Omega)} + \|u_{tt}(t)\| \quad \leq \quad Ct^{-3/2}\|v\|_1 \quad \text{if} \quad v \in \mathbb{H},$$

$$\|u_t(t)\|_{H^2(\Omega)} + \|u_{tt}(t)\| \quad \leq \quad Ct^{-1}\|v\|_2 \quad \text{if} \quad v \in \mathcal{D}(-A) = V \cap H^2(\Omega),$$

$$\|u_t(t)\|_{H^2(\Omega)} + \|u_{tt}(t)\| \quad \leq \quad Ct^{-1/2}\| - Av\|_{H^1(\Omega)} \quad \text{if} \quad v \in \mathcal{D}(-A) \text{ and } -Av \in H^1(\Omega).$$

Using these in the proof of [32, Theorem 5.3] gives the result. ■

Our second preliminary lemma concerns the mild solution SPDE of (5.1).

Lemma 5.10 *Let* X *be the mild solution of (5.1) given in (5.2), let* $0 \leq \gamma < 1$ *and* $t_1, t_2 \in [0, T], \quad t_1 < t_2$.

(i) If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $\|(-A)^{\alpha/2}Q^{1/2}\|_{HS} < \infty$ *with* $0 \leq \alpha \leq 2$ *and suppose* F *satisfies Assumption 5.3 (a). Set* $\sigma = \min(\gamma, 1/2, \alpha/2)$ *then*

$$\mathbf{E}\|X(t_2) - X(t_1)\|^2 \quad \leq \quad C(t_2 - t_1)^{2\sigma}\left(\mathbf{E}\|X_0\|_\gamma^2 + \mathbf{E}\left(\sup_{0\leq s\leq T}(\|F(0)\| + \|X(s)\|)\right)^2 + 1\right).$$

Furthermore

$$\mathbf{E}\|(X(t_2) - O(t_2)) - (X(t_1) - O(t_1))\|^2 \quad \leq \quad C(t_2 - t_1)^{2\gamma} \quad 0 \leq \gamma \leq 1.$$

(ii) If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{(\gamma+1)/2}))$, $\|(-A)^{1/2}Q^{1/2}\|_{HS} < \infty$ *and* $F(X(t)) \in L_2(\mathbb{D}, H^1(\Omega))$, $\forall\, t \in [0, T]$ *with*

$\mathbf{E}\left(\sup_{0\le s\le T}\|F(X(s))\|_{H^1(\Omega)}\right)^2 < \infty$ *then*

$$\mathbf{E}\|X(t_2)-X(t_1)\|^2 \le C(t_2-t_1)^\gamma\left(\mathbf{E}\|X_0\|^2_{(\gamma+1)} + \mathbf{E}\left(\sup_{0\le s\le T}\|F(X(s))\|_{H^1(\Omega)}\right)^2 + 1\right).$$

Proof.

Proof of the first claim of part (i)

Consider the difference

$$\begin{aligned} & X(t_2)-X(t_1) \\ &= (S(t_2)-S(t_1))\,X_0 + \left(\int_0^{t_2} S(t_2-s)F(X(s))ds - \int_0^{t_1} S(t_1-s)F(X(s))ds\right) \\ &\quad + \left(\int_0^{t_2} S(t_2-s)dW(s) - \int_0^{t_1} S(t_1-s)dW(s)\right) \\ &= I + II + III \end{aligned}$$

so that $\mathbf{E}\|X(t_2)-X(t_1)\|^2 \le 3(\mathbf{E}\|I\|^2 + \mathbf{E}\|II\|^2 + \mathbf{E}\|III\|^2)$.

We estimate each of the terms I, II and III. For I, using Proposition 2.6 yields

$$\|I\| = \|S(t_1)(-A)^{-\gamma}(\mathrm{I}-S(t_2-t_1))(-A)^\gamma X_0\| \le C(t_2-t_1)^\gamma\|X_0\|_\gamma.$$

Then $\mathbf{E}\|I\|^2 \le C(t_2-t_1)^{2\gamma}\mathbf{E}\|X_0\|^2_\gamma$. For the term II, we have

$$\begin{aligned} II &= \int_0^{t_1}(S(t_2-s)-S(t_1-s))F(X(s))ds + \int_{t_1}^{t_2} S(t_2-s)F(X(s))ds \\ &= II_1 + II_2. \end{aligned}$$

We now estimate each term II_1 and II_2. For II_1

$$\begin{aligned} \mathbf{E}\|II_1\|^2 &= \mathbf{E}\|\int_0^{t_1}(S(t_2-s)-S(t_1-s))F(X(s))ds\|^2 \\ &\le \mathbf{E}\left(\int_0^{t_1}\|(S(t_2-s)-S(t_1-s))F(X(s))\|ds\right)^2 \\ &\le \left(\int_0^{t_1}\|(S(t_2-s)-S(t_1-s))\|_{L(L^2(\Omega))}ds\right)^2 \mathbf{E}\left(\sup_{0\le s\le T}\|F(X(s))\|\right)^2. \end{aligned}$$

For $0 \leq \gamma < 1$, Proposition 2.6 yields

$$\begin{aligned}
\mathbf{E}\|II_1\|^2 &\leq \left(\int_0^{t_1} \|S(t_1-s)(-A)^{\gamma}(-A)^{-\gamma}(\mathrm{I}-S(t_2-t_1))\|_{L(L^2(\Omega))}ds\right)^2 \mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|\right)^2 \\
&\leq \left(\int_0^{t_1} \|(-A)^{\gamma}S(t_1-s)(-A)^{-\gamma}(\mathrm{I}-S(t_2-t_1))\|_{L(L^2(\Omega))}ds\right)^2 \mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|\right)^2 \\
&\leq C(t_2-t_1)^{2\gamma}\left(\int_0^{t_1}(t_1-s)^{-\gamma}ds\right)^2 \mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|\right)^2 \\
&\leq C(t_2-t_1)^{2\gamma}\,\mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|\right)^2.
\end{aligned}$$

For II_2, using the fact that the semigroup is bounded, we have

$$\begin{aligned}
\mathbf{E}\|II_2\|^2 &= \mathbf{E}\|\int_{t_1}^{t_2} S(t_2-s)F(X(s))ds\|^2 \\
&\leq \mathbf{E}\left(\int_{t_1}^{t_2}\|S(t_2-s)F(X(s))\|ds\right)^2 \\
&\leq \mathbf{E}\left(\int_{t_1}^{t_2}\|F(X(s))\|ds\right)^2 \quad \leq \quad C(t_2-t_1)^2\mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|\right)^2.
\end{aligned}$$

Hence, if F satisfies Assumption 5.3 (a) we have

$$\mathbf{E}\|II\|^2 \leq 2(\mathbf{E}\|II_1\|^2+\mathbf{E}\|II_2\|^2) \leq C(t_2-t_1)^{2\gamma}\mathbf{E}\left(\sup_{0\leq s\leq T}(\|F(0)\|+\|X(s)\|)\right)^2.$$

For term III we have

$$III = \int_0^{t_1}(S(t_2-s)-S(t_1-s))\,dW(s)+\int_{t_1}^{t_2}S(t_2-s)dW(s) \quad = \quad III_1+III_2.$$

First using the Ito isometry property and then $0 \leq \alpha \leq 2$ we have

$$\begin{aligned}
\mathbf{E}\|III_1\|^2 &= \mathbf{E}\|\int_0^{t_1}(S(t_2-s)-S(t_1-s))\,dW(s)\|^2 \\
&= \int_0^{t_1}\mathbf{E}\|\left(S(t_2-s)-S(t_1-s)\right)Q^{1/2}\|_{HS}^2 ds \\
&= \int_0^{t_1}\mathbf{E}\|S(t_2-s)-S(t_1-s))(-A)^{-\alpha/2}(-A)^{\alpha/2}Q^{1/2}\|_{HS}^2 ds.
\end{aligned}$$

Using Proposition 2.6, that $\|(-A)^{\alpha/2}Q^{1/2}\|_{HS} < \infty$ and boundedness of S yields

$$\begin{aligned}
\mathbf{E}\|III_1\|^2 &\leq \int_0^{t_1} \|(-A)^{-\alpha/2}(S(t_2-s)-S(t_1-s))\|^2_{L(L^2(\Omega))}ds \\
&= \int_0^{t_1} \|S(t_1-s)(-A)^{-\alpha/2}(\mathbf{I}-S(t_2-t_1))\|^2_{L(L^2(\Omega))}ds \\
&\leq C\int_0^{t_1} \|(-A)^{-\alpha/2}(\mathrm{I}-S(t_2-t_1))\|^2_{L(L^2(\Omega))}ds \\
&\leq C(t_2-t_1)^{\alpha}.
\end{aligned}$$

Let us estimate $\mathbf{E}\|III_2\|$. Using the Ito isometry again, and that for $0 \leq \alpha \leq 2$ we assume $\|(-A)^{\alpha/2}Q^{1/2}\|_{HS} < \infty$ then $\|Q^{1/2}\|_{HS} < \infty$, with boundedness of S yields

$$\mathbf{E}\|III_2\|^2 = \mathbf{E}\|\int_{t_1}^{t_2} S(t_2-s)dW(s)\|^2 \;=\; \int_{t_1}^{t_2} \|S(t_2-s)Q^{1/2}\|^2_{HS}ds \leq\; C(t_2-t_1).$$

Hence

$$\mathbf{E}\|III\|^2 \leq 2(\mathbf{E}\|III_1\|^2 + \mathbf{E}\|III_2\|^2) \leq C(t_2-t_1)^{2\min(\gamma,\alpha/2,1/2)} = C(t_2-t_1)^{2\sigma}.$$

Combining our estimates of $\mathbf{E}\|I\|^2, \mathbf{E}\|II\|^2$ and $\mathbf{E}\|III\|^2$ ends the first part of the first claim in the lemma.

Proof of the second claim of part (i)

As before consider the difference

$$\begin{aligned}
&(X(t_2)+O(t_2)-(X(t_1)+O(t_1))) \\
&= (S(t_2)-S(t_1))\,X_0 + \left(\int_0^{t_2} S(t_2-s)F(X(s))ds - \int_0^{t_1} S(t_1-s)F(X(s))ds\right) \\
&= I+II,
\end{aligned}$$

so that

$$\mathbf{E}\|(X(t_2)+O(t_2))-(X(t_1)+O(t_1))\|^2 \leq 2(\mathbf{E}\|I\|^2+\mathbf{E}\|II\|^2).$$

We estimate each of the terms I, II . For $0 \leq \gamma \leq 1$, using Proposition 2.6 yields

$$\|I\| \;=\; \|S(t_1)(-A)^{-\gamma}(\mathrm{I}-S(t_2-t_1))(-A)^{\gamma}X_0\| \quad\leq\quad C(t_2-t_1)^{\gamma}\|X_0\|_{\gamma}.$$

Then $\mathbf{E}\|I\|^2 \leq C(t_2 - t_1)^{2\gamma}\mathbf{E}\|X_0\|_\gamma^2$. For the term II, we have

$$\begin{aligned} II &= \int_0^{t_1} (S(t_2 - s) - S(t_1 - s))F(X(s))ds + \int_{t_1}^{t_2} S(t_2 - s)F(X(s))ds \\ &= II_1 + II_2. \end{aligned}$$

We now estimate each term II_1 and II_2. For $\mathbf{E}\|II_2\|^2$ boundedness of S gives

$$\begin{aligned} \mathbf{E}\|II_2\|^2 &\leq \mathbf{E}\left(\int_{t_1}^{t_2} \|S(t_2 - s)F(X(s))\|ds\right)^2 \\ &\leq C\left(t_2 - t_1\right)^2 \mathbf{E}\left(\sup_{0\leq s\leq T} \|F(X(s))\|\right)^2. \end{aligned}$$

For $\mathbf{E}\|II_1\|^2$ we have

$$\begin{aligned} II_1 &= \int_0^{t_1} (S(t_2 - s) - S(t_1 - s))F(X(s))ds \\ &= \int_0^{t_1} (S(t_2 - s) - S(t_1 - s))\left(F(X(s)) - F(X(t_1))\right)ds \\ &+ \int_0^{t_1} (S(t_2 - s) - S(t_1 - s))F(X(t_1))ds \\ &= II_{11} + II_{12}. \end{aligned}$$

Using the Lipschitz condition in Assumption 5.3 (a) with the first claim of (i) yields

$$\begin{aligned} \mathbf{E}\|II_{11}\|^2 &\leq \left(\int_0^{t_1} \|S(t_2 - s) - S(t_1 - s)\|_{L(L^2(\Omega))} \left(\mathbf{E}\|X(s) - X(t_1)\|^2\right)^{1/2} ds\right)^2 \\ &\leq C\left((t_2 - t_1)\int_0^{t_1} (t_1 - s)^{\sigma-1}ds\right)^2 \\ &\leq C\left(t_2 - t_1\right)^2. \end{aligned}$$

Assumption 5.3 (a) gives

$$\begin{aligned} \left(\mathbf{E}\|II_{12}\|^2\right)^{1/2} &\leq \left(\mathbf{E}\|F(X(t_1)\|^2\right)^{1/2} \| \int_0^{t_1} (S(t_2 - s) - S(t_1 - s))ds\|_{L(L^2(\Omega))} \\ &\leq C\| \int_0^{t_1} S(t_2 - s) - S(t_1 - s)ds\|_{L(L^2(\Omega))}. \end{aligned}$$

Using the two transformations $y = t_2 - s,\ y = t_1 - s$ we find

$$\begin{aligned}
\left(\mathbf{E}\|II_{12}\|^2\right)^{1/2} &\leq C\|\int_{t_2-t_1}^{t_2} S(y)dy - \int_0^{t_1} S(y)dy\|_{L(L^2(\Omega))} \\
&\leq C\|\int_{t_2-t_1}^{t_1} S(y)dy + \int_{t_1}^{t_2} S(y)dy - \int_0^{t_1} S(y)dy\|_{L(L^2(\Omega))} \\
&\leq C\|\int_{t_1}^{t_2} S(y)dy - \int_0^{t_2-t_1} S(y)dy\|_{L(L^2(\Omega))} \\
&\leq C(t_2 - t_1).
\end{aligned}$$

Combining the previous estimates ends the proof of the second claim of (i).

Proof of part (ii)

We now prove part (ii) of the lemma. Again let us consider the difference

$$\begin{aligned}
&X(t_2) - X(t_1) \\
&= (S(t_2) - S(t_1))\, X_0 + \left(\int_0^{t_2} S(t_2 - s)F(X(s))ds - \int_0^{t_1} S(t_1 - s)F(X(s))ds\right) \\
&\quad + \left(\int_0^{t_2} S(t_2 - s)dW(s) - \int_0^{t_1} S(t_1 - s)dW(s)\right) \\
&= I + II + III,
\end{aligned}$$

and then

$$\mathbf{E}\|X(t_2) - X(t_1)\|^2_{H^1(\Omega)} \leq 3\left(\mathbf{E}\|I\|^2_{H^1(\Omega)} + \mathbf{E}\|II\|^2_{H^1(\Omega)} + \mathbf{E}\|III\|^2_{H^1(\Omega)}\right).$$

Let us estimate the terms I, II and III and we start with I. If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{(\gamma+1)/2}))$ using Proposition 2.6 yields

$$\begin{aligned}
\|I\|_{H^1(\Omega)} &= \|(-A)^{1/2}S(t_1)(\mathrm{I} - S(t_2 - t_1))X_0\| \\
&= \|(-A)^{1/2}S(t_1)(\mathrm{I} - S(t_2 - t_1))(-A)^{-\gamma/2}(-A)^{\gamma/2}X_0\| \\
&= \|S(t_1)(-A)^{-\gamma/2}(\mathrm{I} - S(t_2 - t_1))(-A)^{(\gamma+1)/2}X_0\| \\
&\leq C(t_2 - t_1)^{\gamma/2}\|X_0\|_{(\gamma+1)}.
\end{aligned}$$

Then

$$\mathbf{E}\|I\|^2_{H^1(\Omega)} \leq C(t_2 - t_1)^\gamma \|X_0\|^2_{(\gamma+1)}.$$

For the term II, we have

$$\begin{aligned} II &= \int_0^{t_1} (S(t_2 - s) - S(t_1 - s))F(X(s))ds + \int_{t_1}^{t_2} S(t_2 - s)F(X(s))ds \\ &= II_1 + II_2. \end{aligned}$$

We now estimate each term above. Using the fact that in $\mathcal{D}((-A)^{1/2})$ we have the equivalency of norm $\|.\|_{H^1(\Omega)} \equiv \|(-A)^{1/2}.\|$, we have

$$\|S(t)\|_{L(H^1(\Omega))} \leq C\|(-A)^{1/2}S(t)\|_{L(L^2(\Omega))} \tag{5.31}$$

where $\|S(t)\|_{L(H^1(\Omega))}$ is the norm of the semigroup viewed as a bounded operator in $H^1(\Omega)$. Indeed using the smoothing properties of the semigroup $S(t)$ in Proposition 2.6 we have

$$v \in \mathcal{D}((-A)^{1/2}) \Rightarrow S(t)v \in \mathcal{D}((-A)^{1/2}),$$

then by the equivalency of norm $\|.\|_{H^1(\Omega)} \equiv \|(-A)^{1/2}.\|$, for $v \in \mathcal{D}((-A)^{1/2})$

$$\|S(t)v\|_{H^1(\Omega)} \leq C\,\|(-A)^{1/2}S(t)v\| \tag{5.32}$$

$$\leq C\,\|(-A)^{1/2}S(t)\|_{L(L^2(\Omega))}\|v\| \tag{5.33}$$

$$\leq C\,\|(-A)^{1/2}S(t)\|_{L(L^2(\Omega))}\|v\|_{H^1(\Omega)}, \tag{5.34}$$

thus

$$\|S(t)\|_{L(H^1(\Omega))} \leq C\|(-A)^{1/2}S(t)\|_{L(L^2(\Omega))}$$

since $\|S(t)\|_{L(H^1(\Omega))}$ is the smallest positive constant L_0 such that

$$\|S(t)v\|_{H^1(\Omega)} \leq L_0\|v\|_{H^1(\Omega)}, \qquad \forall v \in H^1(\Omega).$$

We also have the similar relationship for the operator $S(t_1) - S(t_2)$ with t_1, $t_2 \in [0, T]$.

Using a similar inequality to (5.31) yields

$$\begin{aligned}
\mathbf{E}\|II_1\|^2_{H^1(\Omega)} &= \mathbf{E}\| \int_0^{t_1} (S(t_2-s)-S(t_1-s))F(X(s))ds\|^2_{H^1(\Omega)} \\
&\leq \mathbf{E}\left(\int_0^{t_1} \|(S(t_2-s)-S(t_1-s))F(X(s))\|_{H^1(\Omega)}ds\right)^2 \\
&\leq \left(\int_0^{t_1} \|(S(t_2-s)-S(t_1-s))\|_{L(H^1(\Omega))}ds\right)^2 \mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2 \\
&\leq \left(\int_0^{t_1} \|(-A)^{1/2}(S(t_2-s)-S(t_1-s))\|_{L(L^2(\Omega))}ds\right)^2 \\
&\quad \times \mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2.
\end{aligned}$$

For $\epsilon \in (0,1)$ small enough, we have

$$\begin{aligned}
&\mathbf{E}\|II_1\|^2_{H^1(\Omega)} \\
&\leq \left(\int_0^{t_1} \|(-A)^{(1-\epsilon)/2}S(t_1-s)(-A)^{(\epsilon-1)/2}((\mathrm{I}-S(t_2-t_1)))\|_{L(L^2(\Omega))}ds\right)^2 \\
&\quad \times \mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2 \\
&\leq C(t_2-t_1)^{1-\epsilon}\left(\int_0^{t_1}(t_1-s)^{(1-\epsilon)/2}ds\right)^2 \mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2 \\
&\leq C(t_2-t_1)^{1-\epsilon}\,\mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2.
\end{aligned}$$

We also have using Proposition 2.6

$$\begin{aligned}
\mathbf{E}\|II_2\|^2_{H^1(\Omega)} &= \mathbf{E}\| \int_{t_1}^{t_2} S(t_2-s)F(X(s))ds\|^2_{H^1(\Omega)} \\
&\leq \mathbf{E}\left(\int_{t_1}^{t_2} \|S(t_2-s)F(X(s))\|_{H^1(\Omega)}ds\right)^2 \\
&\leq \mathbf{E}\left(\int_{t_1}^{t_2} \|S(t_2-s)\|_{L(H^1(\Omega))}\|F(X(s))\|_{H^1(\Omega)}ds\right)^2
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{E}\|II_2 &\leq \mathbf{E}\left(\int_{t_1}^{t_2} \|(-A)^{1/2}(S(t_2-s))\|_{L(L^2(\Omega))}\|F(X(s))\|_{H^1(\Omega)}ds\right)^2 \\
&\leq \left(\int_{t_1}^{t_2}(t_2-s)^{-1/2}ds\right)^2 \mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2 \\
&\leq C(t_2-t_1)\mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2.
\end{aligned}$$

Hence, if $F(X(t)) \in L_2(\mathbb{D}, H^1(\Omega)),\ t \in [0,T]$ with $\mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2 < \infty$, we have

$$\mathbf{E}\|II\|^2 \leq 2(\mathbf{E}\|II_1\|^2 + \mathbf{E}\|II_2\|^2) \leq C(t_2-t_1)^{\gamma}\mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2.$$

We also have for the term III

$$\begin{aligned}
III &= \int_0^{t_2} S(t_2-s)dW(s) - \int_0^{t_1} S(t_1-s)dW(s) \\
&= \int_0^{t_1} \left(S(t_2-s)-S(t_1-s)\right)dW(s) + \int_{t_1}^{t_2} S(t_2-s)dW(s) \\
&= III_1 + III_2.
\end{aligned}$$

The Ito isometry property yields

$$\begin{aligned}
\mathbf{E}\|III_1\|^2_{H^1(\Omega)} &= \mathbf{E}\| \int_0^{t_1} \left(S(t_2-s)-S(t_1-s)\right)dW(s)\|^2_{H^1(\Omega)} \\
&\leq \int_0^{t_1} \mathbf{E}\|(-A)^{1/2}\left(S(t_2-s)-S(t_1-s)\right)Q^{1/2}\|^2_{HS}ds \\
&= \int_0^{t_1} \mathbf{E}\|\left(S(t_2-s)-S(t_1-s)\right)(-A)^{1/2}Q^{1/2}\|^2_{HS}ds.
\end{aligned}$$

Using Proposition 2.6, the fact that $S(t)$ is bounded and $\|(-A)^{1/2}Q^{1/2}\|_{HS} < \infty$ yields

$$\begin{aligned}
\mathbf{E}\|III_1\|^2_{H^1(\Omega)} &\leq C\int_0^{t_1} \|(S(t_2-s)-S(t_1-s))\|^2_{L(L^2(\Omega))}ds \\
&= C\int_0^{t_1} \|(-A)^{(1-\epsilon)/2}S(t_1-s)(-A)^{-(1-\epsilon)/2}(\mathrm{I}-S(t_2-t_1))\|^2_{L(L^2(\Omega))}ds \\
&\leq C(t_2-t_1)^{1-\epsilon}\int_0^{t_1}(t_1-s)^{-1+\epsilon}ds \\
&\leq C(t_2-t_1)^{1-\epsilon}.
\end{aligned}$$

with $\epsilon \in (0,1)$ small enough. Let us estimate $\mathbf{E}\|III_2\|^2_{H^1(\Omega)}$. The fact that $\|(-A)^{1/2}Q^{1/2}\|_{HS} < \infty$ yields

$$\begin{aligned}
\mathbf{E}\|III_2\|^2_{H^1(\Omega)} &= \mathbf{E}\| \int_{t_1}^{t_2} S(t_2-s))dW(s)\|^2_{H^1(\Omega)} \\
&\leq \int_{t_1}^{t_2} \|(-A)^{1/2}S(t_2-s)Q^{1/2}\|^2_{HS}ds \\
&= \int_{t_1}^{t_2} \|S(t_2-s)(-A)^{1/2}Q^{1/2}\|^2_{HS}ds \\
&\leq C(t_2-t_1).
\end{aligned}$$

Hence

$$\mathbf{E}\|III\|^2 \leq 2(\mathbf{E}\|III_1\|^2 + \mathbf{E}\|III_2\|^2) \leq C(t_2-t_1)^{\gamma}.$$

Combining the estimates of $\mathbf{E}\|I\|^2, \mathbf{E}\|II\|^2$ and $\mathbf{E}\|III\|^2$ ends the proof. ■

Remark 5.11 *If $\gamma \geq 1$ and with more regularity of the noise ($O(t) \in L_2(\mathbb{D}, \mathcal{D}((-A)^r))$, $r > 1/2$) we have*

$$\boldsymbol{E}\|X(t_2) - X(t_1)\|^2 \leq C(t_2-t_1)^{1-\epsilon}$$

for any $\epsilon \in (0,1)$.

We can prove that we can take $\theta = 1/2$ for $\mathcal{V} = H^1(\Omega)$ or if $O(t) \in L_2(\mathbb{D}, \mathcal{D}(-A))$. We have $\theta \neq 1/2$ and close to $1/2$ if $O(t) \in L_2(\mathbb{D}, \mathcal{D}((-A)^{1/2}))$. These estimates follow those used to estimate III_1 and III_2 in the proof of Lemma 5.10 above.

5.3.2 Proof of Theorem 5.7

We now estimate $\left(\mathbf{E}\|X(t_m) - X^h_m\|^2\right)^{1/2}$. Again we look at the difference between the mild solution and our numerical approximation (5.19). By construction of the approximation

from (5.8) and (5.18) we have that

$$\begin{aligned} X(t_m) - X_m^h &= \overline{X}_{t_m} + O(t_m) - X_m^h \\ &= \overline{X}(t_m) + O(t_m) - \left(Z_m^h + P_h P_N O(t_m)\right) \\ &= \left(\overline{X}(t_m) - Z_m^h\right) + (P_N(O(t_m)) - P_h P_N(O(t_m))) + (O(t_m) - P_N(O(t_m))) \\ &= I + II + III, \end{aligned} \tag{5.35}$$

where $\overline{X}(t)$ is given by (5.9) and Z_m^h by (5.17).

Then $\left(\mathbf{E}\|X(t_m) - X_m^h\|^2\right)^{1/2} \leq (\mathbf{E}\|I\|^2)^{1/2} + (\mathbf{E}\|II\|^2)^{1/2} + (\mathbf{E}\|III\|^2)^{1/2}$ and we estimate each term. Since the first term will require the most work we first estimate the other two.

Let us estimate $(\mathbf{E}\|II\|^2)^{1/2}$. To do this we use the finite element estimate (5.23), regularity of the noise from Assumption 5.5 and the fact that P_N is bounded. Then for $r = 1, 2$ we have

$$\mathbf{E}\|II\|^2 \leq Ch^{2r}\mathbf{E}\|P_N(O(t_m))\|_{H^r(\Omega)} \leq Ch^{2r}\mathbf{E}\|O(t_m)\|_{H^r(\Omega)}.$$

Using $\|.\|_{H^r(\Omega)} \equiv \|(-A)^{r/2}.\|$ in $\mathcal{D}((-A)^{r/2})$, the Ito isometry and the fact that the semigroup is a bounded operator yields

$$\begin{aligned} \mathbf{E}\|II\|^2 &\leq Ch^{2r}\mathbf{E}\|(-A)^{r/2}\int_0^{t_m} S(t_m - s)dW(s)\|^2 \\ &\leq Ch^{2r}\int_0^{t_m} \|(-A)^{r/2}S(t_m - s)\|_{L_2^0}^2 ds \\ &\leq Ch^{2r}\int_0^{T} \|(-A)^{r/2}Q^{1/2}\|_{HS}^2 ds. \end{aligned}$$

Thus, since the noise is in H^r we have $(\mathbf{E}\|II\|^2)^{1/2} \leq Ch^r$.

For the third term III

$$\mathbf{E}\|III\|^2 = \mathbf{E}\|(\mathrm{I} - P_N)O(t_m)\|^2 = \mathbf{E}\|(\mathrm{I} - P_N)(-A)^{-r/2}(-A)^{r/2}O(t_m)\|^2,$$

and so

$$\mathbf{E}\|III\|^2 \leq \|(\mathrm{I} - P_N)(-A)^{-r/2}\|_{L(L^2(\Omega))}^2 \mathbf{E}\|(-A)^{r/2}O(t_m)\|^2 \leq C\left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N} \lambda_j\right)^{-r}.$$

We now turn our attention to the first term $\mathbf{E}\|I\|^2$. Using the definition of $S_{h,\Delta t}$ in (5.16) we can write (5.17) as

$$Z_m^h = S_{h,\Delta t}^m P_h X_0 + \Delta t \sum_{k=0}^{m-1} S_{h,\Delta t}^{(m-k)} P_h F(Z_k^h + P_h P_N O(t_k)).$$

Then using the definition of T_m from (5.25) the first term I can be expanded

$$\begin{aligned}
I &= T_m X_0 + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S(t_m - s)F(X(s)) - S_{h,\Delta t}^{(m-k)} P_h F(Z_k^h + P_h P_N O(t_k))ds \\
&= T_m X_0 + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} P_h (F(X(t_k)) - F(Z_k^h + P_h P_N O(t_k))))ds \\
&\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} P_h (F(X(s)) - F(X(t_k)))ds \\
&\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (S(t_m - t_k) - S_{h,\Delta t}^{(m-k)} P_h) F(X(s))ds \\
&\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (S(t_m - s) - S(t_m - t_k)) F(X(s))ds \\
&= I_1 + I_2 + I_3 + I_4 + I_5. \qquad (5.36)
\end{aligned}$$

Then

$$\left(\mathbf{E}\|I\|^2\right)^{1/2} \le \left(\mathbf{E}\|I_1\|^2\right)^{1/2} + \left(\mathbf{E}\|I_2\|^2\right)^{1/2} + \left(\mathbf{E}\|I_3\|^2\right)^{1/2} + \left(\mathbf{E}\|I_4\|^2\right)^{1/2} + \left(\mathbf{E}\|I_5\|^2\right)^{1/2}.$$

For I_1, if $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $1/2 \le \gamma < 1$, equation (5.26) of Lemma 5.9 gives

$$(\mathbf{E}\|I_1\|^2)^{1/2} \le C(t_m^{-1/2}(h^2 + \Delta t)) \left(\mathbf{E}\|X_0\|_1^2\right)^{1/2}$$

and if $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$, equation (5.27) of Lemma 5.9 gives

$$(\mathbf{E}\|I_1\|^2)^{1/2} \le C(h^2 + \Delta t) \left(\mathbf{E}\|X_0\|_2^2\right)^{1/2}.$$

If F satisfies Assumption 5.3 (a), then using the Lipschitz condition, triangle inequality and the fact that $S_{h,\Delta t}^{(m-k)}$ and P_h are an bounded operators, we have

$$\begin{aligned}
\left(\mathbf{E}\|I_2\|^2\right)^{1/2} &\le C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|F(X(t_k)) - F((Z_k^h + P_h P_N O(t_k))\|^2\right)^{1/2} ds \\
&\le C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2\right)^{1/2} ds.
\end{aligned}$$

Let us estimate $(\mathbf{E}\|I_3\|^2)^{1/2}$. We add in and subtract out $O(s)$ and $O(t_k)$

$$\begin{aligned}
\left(\mathbf{E}\|I_3\|^2\right)^{1/2} &= \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} P_h\left(F(X(s)) - F(X(t_k))\right) ds\|^2\right)^{1/2} \\
&\leq \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} P_h\left(F(X(s)) - F(X(t_k)+O(s)-O(t_k))\right) ds\|^2\right)^{1/2} \\
&\quad + \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} P_h\left(F(X(t_k)+O(s)-O(t_k)) - F(X(t_k))\right) ds\|^2\right)^{1/2} \\
&:= \left(\mathbf{E}\|I_3^1\|^2\right)^{1/2} + \mathbf{E}\left(\|I_3^2\|^2\right)^{1/2}.
\end{aligned}$$

Applying the Lipschitz condition in Assumption 5.5, using the fact the semigroup is bounded and according to Lemma 5.10, for $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma)), \quad 0 \leq \gamma \leq 1$ we therefore have

$$\begin{aligned}
\left(\mathbf{E}\|I_3^1\|^2\right)^{1/2} &\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|(X(s)-O(s)) - (X(t_k)-O(t_k))\|^2\right)^{1/2} ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (s-t_k)^\gamma ds \quad \leq \quad C\Delta t^\gamma.
\end{aligned}$$

Let us now estimate $\mathbf{E}\left(\|I_3^2\|^2\right)^{1/2}$. The analysis below follows the same steps as in [89] although the approximating semigroup $S_{h,\Delta t}$ is different here. Applying a Taylor expansion to F gives

$$\mathbf{E}\left(\|I_3^2\|^2\right)^{1/2} \leq I_3^{21} + I_3^{22} + I_3^{23},$$

with

$$\begin{aligned}
I_3^{21} &= \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} P_h F'(X(t_k))(O(s) - S(s-t_k)O(t_k)) ds\|^2\right)^{1/2} \\
I_3^{22} &= \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} P_h F'(X(t_k))(S(s-t_k)O(t_k) - O(t_k)) ds\|^2\right)^{1/2} \\
I_3^{23} &= \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} \int_0^1 G(1-r) dr ds\|^2\right)^{1/2},
\end{aligned}$$

where G is the expression $G := P_h F''(X(t_k)) + r(O(s) - O(t_k))(O(s) - O(t_k), O(s) - O(t_k))$. Using the fact that $O(t_2) - S(t_2 - t_1)O(t_1)$, $0 \le t_1 < t_2 \le T$ is independent of $\mathcal{F}_{t_1}$, one can show, as in [89], that

$$\left(I_3^{21}\right)^2 = \sum_{k=0}^{m-1} \mathbf{E}\| \int_{t_k}^{t_{k+1}} S_{h,\Delta t}^{(m-k)} P_h F'(X(t_k))(O(s) - S(s - t_k)O(t_k))ds\|^2.$$

Therefore as $S_{h,\Delta t}$ is bounded we have

$$\begin{aligned} I_3^{21} &\le \left(\sum_{k=0}^{m-1} \left(\int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|S_{h,\Delta t}^{(m-k)} P_h F'(X(t_k))(O(s) - S(s - t_k)O(t_k))\|^2\right)^{1/2} ds\right)^2\right)^{1/2} \\ &\le C\left(\sum_{k=0}^{m-1} \left(\int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|P_h F'(X(t_k))(O(s) - S(s - t_k)O(t_k))\|^2\right)^{1/2} ds\right)^2\right)^{1/2}. \end{aligned}$$

By using Hölder's inequality, the following inequality holds

$$\left(\int_a^b f(x)dx\right)^2 \le (b - a)\int_a^b f(x)^2 dx, \tag{5.37}$$

by assuming that f and f^2 are integrable in the bounded interval $[a, b]$. Using (5.37) with Assumption 5.5, Assumption 5.3(a) and Proposition 2.6 yields

$$\begin{aligned} I_3^{21} &\le C\Delta t^{1/2}\left(\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \mathbf{E}\|P_h F'(X(t_k))(O(s) - S(s - t_k)O(t_k))\|^2 ds\right)^{1/2} \\ &\le C\Delta t^{1/2}\left(\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \mathbf{E}\|(O(s) - S(s - t_k)O(t_k))\|^2 ds\right)^{1/2} \\ &\le C\Delta t^{1/2}\left(\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\left(\mathbf{E}\|O(s) - O(t_k)\|^2\right)^{1/2} + \left(\mathbf{E}\|(S(s - t_k) - \mathbf{I})O(t_k))\|^2\right)^{1/2}\right)^2 ds\right)^{1/2} \end{aligned}$$

$$\begin{aligned} I_3^{21} &\le C\Delta t^{1/2}\left(\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left((s - t_k)^\theta + (s - t_k)^{r/2}\left(\mathbf{E}\|O(t_k)\|_r^2\right)^{1/2}\right)^2 ds\right)^{1/2} \\ &\le C\Delta t^{1/2+\theta}. \end{aligned}$$

Let us estimate I_3^{22}.

$$\begin{aligned} I_3^{22} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E} \| S_{h,\Delta t}^{(m-k)} P_h (-A)^{1/2} (-A)^{-1/2} F'(X(t_k)) (S(s-t_k) - \mathbf{I}) O(t_k)) \|^2 \right)^{1/2} ds \\ &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \| S_{h,\Delta t}^{(m-k)} P_h (-A)^{1/2} \|_{L(L^2(\Omega))} \left(\mathbf{E} \| (-A)^{-1/2} F'(X(t_k)) (S(s-t_k) - \mathbf{I}) O(t_k)) \|^2 \right)^{1/2} ds. \end{aligned}$$

Since $P_h(-A)^{1/2} = (-A_h)^{1/2}$ and $S_{h,\Delta t}$ satisfies the smoothing properties analogous to $S(t)$ independently of h (see for example [32, 88]), and in particular

$$\| S_{h,\Delta t}^m (-A_h)^{1/2} \|_{L(L^2(\Omega))} = \| (-A_h)^{1/2} S_{h,\Delta t}^m \|_{L(L^2(\Omega))} \leq C t_m^{-1/2}, \quad t_m = m\Delta t > 0,$$

we therefore have

$$I_3^{22} \leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - t_k)^{-1/2} \left(\mathbf{E} \| (-A)^{-1/2} F'(X(t_k)) ((S(s-t_k) - \mathbf{I}) O(t_k)) \|^2 \right)^{1/2} ds.$$

The usual identification of $H = L^2(\Omega)$ to its dual yields

$$\begin{aligned} I_3^{22} &\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - t_k)^{-1/2} \\ &\quad \times \left(\mathbf{E} \left(\sup_{\|v\| \leq 1} |\langle v, (-A)^{-1/2} F'(X(t_k)) ((S(s-t_k) - \mathbf{I}) O(t_k)) \rangle| \right)^2 \right)^{1/2} ds, \end{aligned}$$

where $\langle , \rangle = (,)$ and we change the notation merely to emphasize that H is identified to its dual space. The fact that $(-A)^{-1/2}$ is self-adjoint implies that $\left((-A)^{-1/2} F'(X)\right)^* = F'(X)^* (-A)^{-1/2}$. This combined with the fact that $\mathcal{D}((-A)^{1/2}) \subset H$ thus $H = H^* \subset \mathcal{D}((-A)^{-1/2}) = \left(\mathcal{D}((-A)^{1/2})\right)^*$ continuously and Assumption 5.3 yields

$$
\begin{aligned}
I_3^{22} &\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m - t_k)^{-1/2} \\
&\quad \times \left(\mathbf{E}\left(\sup_{\|v\|\leq 1} |\langle F'(X(t_k))^*(-A)^{-1/2}v, (S(s-t_k)-\mathbf{I})O(t_k)\rangle| \right)^2 \right)^{1/2} ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m - t_k)^{-1/2} \\
&\quad \times \left(\mathbf{E}\left(\sup_{\|v\|\leq 1} \|F'(X(t_k))^*(-A)^{-1/2}v\|_1 \|(S(s-t_k)-\mathbf{I})O(t_k))\|_{-1} \right)^2 \right)^{1/2} ds.
\end{aligned}
$$

We also have

$$
\begin{aligned}
I_3^{22} &\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m - t_k)^{-1/2} \left(\mathbf{E}\left(1+\|X(t_k)\|_1\right)^2 \| \left(S(s-t_k)-\mathbf{I})O(t_k)\|_{-1}\right)^2 \right)^{1/2} ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m - t_k)^{-1/2} \left(\mathbf{E}\left(1+\|X(t_k)\|_1\right)^4 \right)^{1/4} \left(\mathbf{E}\left(\|S(s-t_k)-\mathbf{I})O(t_k)\|_{-1}\right)^4 \right)^{1/4} ds. \\
&\leq C\sum_{k=0}^{m-1} (t_m - t_k)^{-1/2} \left(1+\left(\mathbf{E}\|X(t_k)\|_1^4\right)^{1/4}\right) \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\left(\|S(s-t_k)-\mathbf{I})O(t_k)\|_{-1}\right)^4 \right)^{1/4} ds \\
&\leq C\sum_{k=0}^{m-1} (t_m - t_k)^{-1/2} \int_{t_k}^{t_{k+1}} \|(-A)^{-(r/2+1/2)}\left(S(s-t_k)-\mathbf{I}\right)\|_{L(L^2(\Omega))} \left(\mathbf{E}\|O(t_k)\|_r^4\right)^{1/4} ds \\
&\leq C\sum_{k=0}^{m-1} (t_m - t_k)^{-1/2} \int_{t_k}^{t_{k+1}} \|(-A)^{1/2-r/2}(-A)^{-1}(S(s-t_k)-\mathbf{I})\|_{L(L^2(\Omega))} ds.
\end{aligned}
$$

Using Proposition 2.6 and the fact that $(-A)^{1/2-r/2}$ is bounded as $r = 1, 2$ yields

$$
I_3^{22} \leq C\sum_{k=0}^{m-1} (t_m - t_k)^{-1/2} \int_{t_k}^{t_{k+1}} (s-t_k)\, ds = C\Delta t^{3/2}\sum_{k=0}^{m-1} (m-k)^{-1/2}.
$$

As the sum above can be bounded by $2M^{1/2}$ we have

$$
I_3^{21} + I_3^{22} \leq C(\Delta t + \Delta t^{1/2+\theta}) \leq C(\Delta t^{2\theta}).
$$

Let us estimate I_3^{23}. Using the fact that $S_{h,\Delta t}^{(m-k)}$ is bounded for any m, k with Assumption 5.3 and Assumption 5.5 yields (with $G = P_h F''(X(t_k) + r(O(s) - O(t_k)))(O(s) - O(t_k), O(s) - O(t_k))$)

$$
\begin{aligned}
I_3^{23} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \|S_{h,\Delta t}^{(m-k)}\|_{L(L^2(\Omega))} \int_0^1 \left(\mathbf{E}\|G\|^2\right)^{1/2} dr ds \\
&\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \int_0^1 \left(\mathbf{E}\|O(s) - O(t_k)\|_{\mathcal{V}}^4\right)^{1/2} dr ds \\
&\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\left(\mathbf{E}\|O(s) - O(t_k)\|_{\mathcal{V}}^4\right)^{1/4}\right)^2 ds \\
&\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (s - t_k)^{2\theta} \, ds \leq C(\Delta t)^{2\theta}.
\end{aligned}
$$

Combining $I_3^{21} + I_3^{22}$ and I_3^{23} yields the following estimation

$$
\mathbf{E}\left(\|I_3\|^2\right)^{1/2} \leq C(\Delta t^{2\theta}) \leq C(\Delta t^{\sigma}).
$$

We now estimate I_4. Using equation (5.26) of Lemma 5.9, if $F(X(t)) \in L_2(\mathbb{D}, \mathbb{H})$ with $\sup_{0 \leq s \leq T} \mathbf{E}\|F(X(s))\|_1^2 < \infty$ yields

$$
\begin{aligned}
\left(\mathbf{E}\|I_4\|^2\right)^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|T_{m-k} F(X(s))\|_1^2\right)^{1/2} ds \\
&\leq C(h^2 \sup_{0 \leq s \leq T} \left(\mathbf{E}\|F(X(s))\|_1^2\right)^{1/2} \left(\Delta t \sum_{k=0}^{m-1} t_{m-k}^{-1/2}\right).
\end{aligned}
$$

Note that we can bound $\Delta t \sum_{k=0}^{m-1} t_{m-k}^{-1/2}$ by $2\sqrt{T}$ then

$$
\left(\mathbf{E}\|I_4\|^2\right)^{1/2} \leq C(h^2 + \Delta t) \left(\sup_{0 \leq s \leq T} \mathbf{E}\|F(X(s))\|_1^2\right)^{1/2}.
$$

Finally we estimate $\left(\mathbf{E}\|I_5\|^2\right)^{1/2}$. Using Proposition 2.6, we have for $0 \leq t_1 < t_2 \leq T$

$$
\|S(t_2) - S(t_1)\|_{L(L^2(\Omega))} = \|(-A)S(t_1)(-A)^{-1}\left(\mathbf{I} - S(t_2 - t_1)\right)\|_{L(L^2(\Omega))} \leq \frac{(t_2 - t_1)}{t_1},
$$

then

$$\begin{aligned}
\left(\mathbf{E}\|I_5\|^2\right)^{1/2} &\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \|S(t_m-s)-S(t_m-t_k)\|_{L(L^2(\Omega))}ds \left(\sup_{0\leq s\leq T}\mathbf{E}\|F(X(s))\|^2\right)^{1/2} \\
&\leq C\left(\Delta t + \sum_{k=0}^{m-2}\int_{t_k}^{t_{k+1}}\left(\frac{s-t_k}{t_m-s}\right)ds\right) \\
&\leq C\left(\Delta t + \sum_{k=0}^{m-2}\left((m-k-1)\Delta t\right)^{-1}\int_{t_k}^{t_{k+1}}(s-t_k)\,ds\right) \\
&\leq C\left(\Delta t + \Delta t\sum_{k=0}^{m-2}(m-k-1)^{-1}\right).
\end{aligned}$$

Noting that the sum above is bounded by $\ln(M)$ we have

$$\left(\mathbf{E}\|I_5\|^2\right)^{1/2} \leq C(\Delta t + \Delta t|\ln(\Delta t)|).$$

For $F(X(t)) \in L_2(\mathbb{D},\mathbb{H})$ with $\sup_{0\leq s\leq T}\mathbf{E}\|F(X(s)\|_1^2 < \infty$, combining the previous estimates for the term I yields: If $X_0 \in L_2(\mathbb{D},\mathcal{D}(-A))$,

$$\left(\mathbf{E}\|I\|^2\right)^{1/2} \leq C\left(h^2+\Delta t^\sigma + \Delta t|\ln(\Delta t)| + \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}\left(\mathbf{E}\|X(t_k)-X_k^h\|^2\right)^{1/2}ds\right).$$

If $X_0 \in L_2(\mathbb{D},\mathcal{D}((-A)^\gamma))$

$$\left(\mathbf{E}\|I\|^2\right)^{1/2} \leq C\left(t_m^{-1/2}(h^2+\Delta t^\sigma) + \Delta t|\ln(\Delta t)| + \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}\left(\mathbf{E}\|X(t_k)-X_k^h\|^2\right)^{1/2}ds\right).$$

Finally we combine all our estimates on I, II and III to get $\left(\mathbf{E}\|I\|^2\right)^{1/2}, \left(\mathbf{E}\|II\|^2\right)^{1/2}$ and $\left(\mathbf{E}\|III\|^2\right)^{1/2}$ and use the discrete Gronwall lemma to complete the proof.

5.3.3 Proof of Theorem 5.8

We now prove convergence in $H^1(\Omega)$ and estimate $\left(\mathbf{E}\|X(t_m)-X_m^h\|^2\right)^{1/2}$. For the proof we follow the same steps as in previous section for Theorem 5.7. We now estimate (5.35) in the H^1 norm.

The estimates of the terms II and III follow as in Section 5.3.2 and we find

$$\left(\mathbf{E}\|II\|^2_{H^1(\Omega)}\right)^{1/2} \leq Ch \qquad \text{and} \qquad \left(\mathbf{E}\|III\|^2_{H^1(\Omega)}\right)^{1/2} \leq C\left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-1/2}.$$

We concentrate instead on estimating the first term $I = I_1 + I_2 + I_3 + I_4 + I_5$ in (6.14) and estimates on I_1 follow immediately from Lemma 5.9.

If F satisfies Assumption 5.3 (b), then using the Lipschitz condition, the triangle inequality, the fact that P_h is an bounded operator and $S_{h,\Delta t}$ satisfies the smoothing property analogous to $S(t)$ independently of h [32], ie

$$\|S^m_{h,\Delta t} v\|^2_{H^1(\Omega)} \leq C t_m^{-1/2} \|v\| \qquad v \in V_h \qquad t_m > 0,$$

we have

$$\begin{aligned}
(\mathbf{E}\|I_2\|^2_{H^1(\Omega)})^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (\mathbf{E}\|S^{(m-k)}_{h,\Delta t} P_h (F(X(t_k)) - F((Z^h_k + P_h P_N O(t_k))))\|^2_{H^1(\Omega)})^{1/2} ds \\
&\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} t_{m-k}^{-1/2} (\mathbf{E}\|F(X(t_k)) - F((Z^h_k + P_h P_N O(t_k))\|^2)^{1/2} ds \\
&\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} t_{m-k}^{-1/2} (\mathbf{E}\|X(t_k) - X^h_k\|^2_{H^1(\Omega)})^{1/2} ds.
\end{aligned}$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$, again using Lipschitz condition, triangle inequality, smoothing property of $S_{h,\Delta t}$, but with Lemma 5.10 gives

$$\begin{aligned}
(\mathbf{E}\|I_3\|^2_{H^1(\Omega)})^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (\mathbf{E}\|S^{(m-k)}_{h,\Delta t} P_h (F(X(s)) - F(X(t_k))\|^2_{H^1(\Omega)})^{1/2} ds \\
&\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} t_{m-k}^{-1/2} (\mathbf{E}\|F(X(s)) - F(X(t_k))\|)^{1/2} ds \\
&\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} t_{m-k}^{-1/2} (\mathbf{E}\|X(s) - X(t_k)\|^2_{H^1(\Omega)})^{1/2} ds \\
&\leq C \left(\sum_{k=0}^{m-1} t_{m-k}^{-1/2} \int_{t_k}^{t_{k+1}} (s - t_k)^{(1-2\epsilon)/2} ds \right) \\
&\quad \times \left(\mathbf{E}\|X_0\|_2^2 + \left(\mathbf{E} \sup_{0 \leq s \leq T} \|F(X(s))\|_{H^1(\Omega)} \right)^2 + 1 \right) \\
&\leq C(\Delta t)^{(1/2-\epsilon)} \left(\Delta t \sum_{k=0}^{m-1} t_{m-k}^{-1/2} \right) \left(\mathbf{E}\|X_0\|_2^2 + \left(\mathbf{E} \sup_{0 \leq s \leq T} \|F(X(s))\|_{H^1(\Omega)} \right)^2 + 1 \right),
\end{aligned}$$

$\epsilon \in (0, 1/4)$ small enough.

As in the previous theorem, we use the fact that $\Delta t \sum_{k=0}^{m-1} t_{m-k}^{-1/2} \leq 2\sqrt{T}$.

For $F(X(t)) \in L_2(\mathbb{D}, \mathbb{H})$ with $\sup_{0\leq s\leq T} \mathbf{E}\|F(X(s))\|_1^2 < \infty$, then by (5.28) of Lemma 5.9 we find

$$\begin{aligned}(\mathbf{E}\|I_4\|^2_{H^1(\Omega)})^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (\mathbf{E}\|T_{m-k}F(X(s))\|^2_{H^1(\Omega)})^{1/2} ds \\ &\leq C\Delta t \left(\sum_{k=0}^{m-1} t_{m-k}^{-1/2} h + t_{m-k}^{-1} \Delta t \right) \left(\sup_{0\leq s\leq T} \mathbf{E}\|F(X(s))\|^2_{H^1(\Omega)} \right)^{1/2}.\end{aligned}$$

Note that $\Delta t \sum_{k=0}^{m-1} t_{m-k}^{-1} \leq \ln(T/\Delta t)$ to get

$$\begin{aligned}\left(\mathbf{E}\|I_4\|^2_{H^1(\Omega)}\right)^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|T_{m-k}F(X(s))\|^2_{H^1(\Omega)}\right)^{1/2} ds \\ &\leq C(h + \Delta t \ln(T/\Delta t)) \left(\sup_{0\leq s\leq T} \mathbf{E}\|F(X(s))\|^2_{H^1(\Omega)} \right)^{1/2}.\end{aligned}$$

Finally, using the equivalency $\|.\|_{H^1(\Omega)} \equiv \|(-A)^{1/2}.\|$ in $\mathcal{D}((-A)^{1/2})$, we obviously have for $0 \leq t_1 < t_2 \leq T$

$$\|S(t_2) - S(t_1)\|_{L(H^1(\Omega))} \leq C\|(-A)^{3/2} S(t_1)(-A)^{-1} \left(\mathbf{I} - S(t_2 - t_1)\right)\|_{L(L^2(\Omega))} \leq C\frac{(t_2 - t_1)}{t_1^{3/2}}$$

so with splitting yields

$$\begin{aligned}(\mathbf{E}\|I_5\|^2_{H^1(\Omega)})^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \|S(t_m - s) - S(t_m - t_k)\|_{L(H^1(\Omega))} ds \left(\sup_{0\leq s\leq T} \mathbf{E}\|F(X(s)\|^2_{H^1(\Omega)} \right)^{1/2} \\ &\leq C \left(\int_{t_{m-1}}^{t_m} \|S(t_m - s) - S(t_m - t_{m-1})\|_{L(H^1(\Omega))} ds \right. \\ &\quad \left. + \sum_{k=0}^{m-2} \int_{t_k}^{t_{k+1}} \left(\frac{s - t_k}{(t_m - s)^{3/2}} \right) ds \right)\end{aligned}$$

$$
\begin{aligned}
(\mathbf{E}\|I_5\|^2_{H^1(\Omega)})^{1/2} &\leq C\left(\int_{t_{m-1}}^{t_m}\left((t_m-s)^{-1/2}+\Delta t^{-1/2}\right)ds\right.\\
&\qquad \left.+\sum_{k=0}^{m-2}(t_m-t_k-\Delta t)^{-3/2}\int_{t_k}^{t_{k+1}}(s-t_k)ds\right)\\
&\leq C\left(\Delta t^{1/2}+\Delta t^{1/2}\sum_{k=0}^{m-2}(m-k-1)^{-3/2}\right).
\end{aligned}
$$

Since the sum above can be bounded by 2 we have that $(\mathbf{E}\|I_5\|^2)^{1/2} \leq C\Delta t^{1/2}$.

Combining our estimates, and using that $\Delta t^{(1-\gamma/2)}\ln(T/\Delta t)$ is bounded as $\Delta t \to 0$, for $F(X(t)) \in L_2(\mathbb{D},\mathbb{H})$ with $\sup_{0\leq s\leq T}\mathbf{E}\|F(X(s))\|_1^2 < \infty$, we have that : If $X_0 \in L_2(\mathbb{D},\mathcal{D}(-A))$ then

$$
(\mathbf{E}\|I\|^2_{H^1(\Omega)})^{1/2} \leq C\left((h+\Delta t^{1/2-\epsilon}t_m^{-1/2})+\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} t_{m-k}^{-1/2}(\mathbf{E}\|X(t_k)-X_k^h\|^2_{H^1(\Omega)})^{1/2}ds\right).
$$

If $X_0 \in L_2(\mathbb{D},\mathcal{D}(-A))$ with $-AX_0 \in L_2(\mathbb{D},H^1(\Omega))$ then

$$
(\mathbf{E}\|I\|^2_{H^1(\Omega)})^{1/2} \leq C\left((h+\Delta t^{1/2-\epsilon})+\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} t_{m-k}^{-1/2}(\mathbf{E}\|X(t_k)-X_k^h\|^2_{H^1(\Omega)})^{1/2}ds\right).
$$

where $C>0$ depending of the T, the initial solution X_0, the mild solution X, the nonlinear function F.

Combining our estimates $(\mathbf{E}\|I\|^2)^{1/2}, (\mathbf{E}\|II\|^2)^{1/2}$ and $(\mathbf{E}\|III\|^2)^{1/2}$ and using the discrete Gronwall lemma concludes the proof.

5.4 Numerical Simulations

5.4.1 A linear reaction–diffusion equation

As a simple example consider the reaction diffusion equation

$$
dX = (D\Delta X - 0.5X)dt + dW \qquad \text{given} \quad X(0)=X_0,
$$

in the time interval $[0,T]$ with diffusion coefficient $D=1/100$ and homogeneous Neumann boundary conditions on the domain $\Omega=[0,L_1]\times[0,L_2]$. Notice that $A=D\Delta$ does not satisfy Assumption 5.2 as 0 is an eigenvalue. During the simulations we need to manage the

singularity of (5.40) at $\lambda_0 = 0$ or use the perturbed operator $A = D\Delta + \epsilon\mathbf{I}$, $\epsilon > 0$. We take $L_1 = L_2 = 1$. Our function $F(u) = 0.5u$ is linear and obviously satisfies Assumption 5.3 (a). In general we are interested in nonlinear F however for this linear system we can find an exact solution to compare our numerics to. The eigenfunctions $\{e_i^{(1)} e_j^{(2)}\}_{i,j\geq 0}$ of the operator $A = \Delta$ here are given by

$$e_0^{(l)} = \sqrt{\frac{1}{L_l}}, \quad \lambda_0^{(l)} = 0, \quad e_i^{(l)} = \sqrt{\frac{2}{L_l}} \cos(\lambda_i^{(l)} x), \quad \lambda_i^{(l)} = \frac{i\,\pi}{L_l}$$

where $l \in \{1, 2\}$ and $i = 1, 2, 3, \cdots$ with the corresponding eigenvalues $\{\lambda_{i,j}\}_{i,j\geq 0}$ given by $\lambda_{i,j} = (\lambda_i^{(1)})^2 + (\lambda_j^{(2)})^2$. Recall that we assumed for convenience that the eigenfunctions of covariance operator Q and A are the same.

We can relate the spatial regularity of the noise to the spatial correlation and eigenvalues q_i of the covariance operator that appear in the representation (4.3). For $O(t) \in L_2(\mathbb{D}, H^r(\Omega)),\ r = 1, 2$, this corresponds to polynomial decay in q_i as in [90, 91] so that

$$q_{i,j} = \Gamma\,(i + j)^{-r/2}\,, \qquad r > 0. \tag{5.38}$$

We also consider exponential correlation as in [75, 81, 92] where

$$\mathbf{E}W((x_1, y_1), t), W((x_2, y_2), t') = C_r((x_1, y_1), (x_2, y_2)) \min(t, t')$$

and

$$C_r((x_1, y_1); (x_2, y_2)) = \frac{\Gamma}{4b_1 b_2} \exp\left(-\frac{\pi}{4}\left[\frac{(x_2 - x_1)^2}{b_1^2} + \frac{(y_2 - y_1)^2}{b_2^2}\right]\right)$$

where b_1, b_2 are spatial correlation lengths in x and y and $\Gamma > 0$.

It is well known [19, 93] that if b and λ are two real numbers the following result holds

$$\int_{-\infty}^{+\infty} \exp\left(-\frac{\pi}{4}\left(\frac{x^2}{b^2}\right)\right) \cos(\lambda x) dx = 2b \exp\left[-\frac{1}{\pi}\,(\lambda b)^2\right]. \tag{5.39}$$

Let us put the noise W in form of the representation (4.3). Recall [26] that the covariance operator Q may be defined for $f \in L^2(\Omega)$ by

$$Qf(x) = \int_\Omega C_r(x, y) f(y) dy.$$

Indeed we have

$$
\begin{aligned}
& 4b_1b_2\int_0^{L_1}\int_0^{L_2} C_r((x_1,y_1);(x_2,y_2))\cos(\lambda_i^{(1)}x_2)\cos(\lambda_j^{(2)}y_2)dy_2dx_2 \\
&= \Gamma\int_0^{L_1}\exp\left(-\frac{\pi}{4}\left(\frac{(x_2-x_1)^2}{b_1^2}\right)\right)\cos(\lambda_i^{(1)}x_2)dx_2 \\
&\quad\times\int_0^{L_2}\exp\left(-\frac{\pi}{4}\left[\frac{(y_2-y_1)^2}{b_2^2}\right]\right)\cos(\lambda_j^{(2)}y_2)dy_2 \\
&= \Gamma\int_{-x_1}^{L_1-x_1}\exp\left(-\frac{\pi}{4}\left(\frac{x^2}{b_1^2}\right)\right)\cos(\lambda_i^{(1)}(x+x_1))dx \\
&\quad\times\int_{-y_1}^{L_2-y_1}\exp\left(-\frac{\pi}{4}\left(\frac{x^2}{b_2^2}\right)\right)\cos(\lambda_j^{(2)}(x+y_1)dx).
\end{aligned}
$$

For $b_i \ll L_i$, because of the strong decay, we approximate the integral in the finite domain by the integral in infinite domain where we can evaluate exactly

$$
\begin{aligned}
& 4b_1b_2\int_0^{L_1}\int_0^{L_2} C_r((x_1,y_1);(x_2,y_2))\cos(\lambda_i^{(1)}x_2)\cos(\lambda_j^{(2)}y_2)dy_2dx_2 \\
&\approx \Gamma\int_{-\infty}^{+\infty}\exp\left(-\frac{\pi}{4}\left(\frac{x^2}{b_1^2}\right)\right)\cos(\lambda_i^{(1)}(x+x_1))dx \\
&\quad\times\int_{-\infty}^{+\infty}\exp\left(-\frac{\pi}{4}\left(\frac{x^2}{b_2^2}\right)\right)\cos(\lambda_j^{(2)}(x+y_1))dx \\
&= 4b_1b_2\cos(\lambda_i^{(1)}x_1)\cos(\lambda_j^{(2)}y_1)\,\Gamma\exp\left(-\frac{1}{\pi}\left((\lambda_i^{(1)}b_1)^2+(\lambda_j^{(2)}b_2)^2\right)\right).
\end{aligned}
$$

It is important to notice that in the previous expressions we have used the fact that

$$
\int_{-\infty}^{+\infty}\exp\left(-\frac{\pi}{4}\left(\frac{x^2}{b_i^2}\right)\right)\cos(\lambda_j^{(i)}x)dx = 2b_i\exp\left[-\frac{1}{\pi}\left((\lambda_j^{(i)}b_i)^2\right)\right] \quad i\in\{1,2\}
$$

by (5.39) and

$$
\int_{-\infty}^{+\infty}\exp\left(-\frac{\pi}{4}\left(\frac{x^2}{b_i^2}\right)\right)\sin(\lambda_j^{(i)}x)dx = 0
$$

because the integrand is an odd function. Then the corresponding values of $\{q_{i,j}\}_{i+j>0}$ in the representation (4.3) is given by

$$
q_{i,j} = \Gamma\exp\left[-\frac{1}{2\pi}\left((\lambda_i^{(1)}b_1)^2+(\lambda_j^{(2)}b_2)^2\right)\right].
$$

In the implementation of our modified scheme at every time step $O(t_{k+1})$ is generated using $O(t_k)$ from the following relation

$$O(t_{k+1}) = e^{A\Delta t}O(t_k) + \int_{t_k}^{t_{k+1}} e^{(t_{k+1}-s)A}dW(s),$$

where $O(0) = 0$. We expand in Fourier space and apply the Ito isometry in each mode and project onto N modes to obtain for $k = 1, 2, \ldots, M-1$

$$(e_i, O(t_{k+1})) = e^{-\lambda_i \Delta t}(e_i, O(t_k)) + \left(\frac{q_i}{2\lambda_i}\left(1 - e^{-2\lambda_i \Delta t}\right)\right)^{1/2} R_{i,k}, \tag{5.40}$$

where $R_{i,k}$ are independent, standard normally distributed random variables with means 0 and variance 1, and $i \in \mathcal{I}_N = \{1, 2, 3, ..., N\}^2$. These are the linear functionals used by Jentzen and Kloeden in [82]. The noise is then projected onto the finite element space by P_h. We require values of the noise at the nodes in the finite element. If we evaluate $O(x, t)$ at these mesh points the projection P_h becomes trivial.

In our simulations we examined both a finite element and a finite volume discretization in space. For the finite element discretization we take $\Delta x = \Delta y = 1/100$ for H^r noise, $r = 1, 2$ and $\Delta x = \Delta y = 1/220$, $b_1 = b_2 = 0.02$, $\Gamma = 0.01$ for noise with an exponential covariance. The finite element triangulation was constructed so that the center of the control volume for the finite volume method was a vertex in finite element mesh.

In Figure 5.1 (a) we take noise in H^r in space with $r = 1, 2$ (i.e. $O(t) \in L_2(\mathbb{D}, H^r(\Omega))$) and examine the root mean square $L^2(\Omega)$ error as we change the step size. We denote the finite element discretization using the modified scheme with a linear functional of the noise 'modifiedImplicitfem$r$', $r = 1, 2$ on the graph. For the finite volume discretization we have implemented both the new modified method "modifiedImplicitfvm$r$", $r = 1, 2$ and a standard implicit Euler–Maruyama method denoted "Implicitfvm$r$", $r = 1, 2$. We see that the observed rate of convergence for the finite element discretization agrees with theorem Theorem 5.7. We also observe that the error decreases as the regularity increases from $r = 1$ to $r = 2$. More importantly we see that the error using the new modified scheme is better in all the cases than using the standard scheme and the rate of convergence is approximately 1 for $r = 1, 2$. Indeed we observe numerically a slower rate of convergence for the the standard scheme of 0.5 and 0.6 for $r = 1$ and $r = 2$ respectively.

In Figure 5.1 (b) we show results with the exponential covariance function, as the noise is certainly in H^r, $r = 1$ or 2 we expect a rate of convergence close to one. The figure com-

pares the finite element discretization ("modifiedimplicitfem") with the modified scheme against the finite volume ("modifiedimplicitfvm") discretization with the new scheme and the standard scheme ("implicitfvm"). We again see the improved accuracy in the modified scheme with a convergence rate of close to one and that the finite element and finite volume discretization are of the same order. Here convergence for the standard scheme is numerically of order 0.75. For a large value of Γ the order of convergence of the standard semi implicit scheme reduces, but we still have the same order of convergence for the modified scheme (see Section 6.4.1).

5.4.2 Stochastic advection diffusion reaction

As a more challenging example we consider the stochastic advection diffusion reaction SPDE

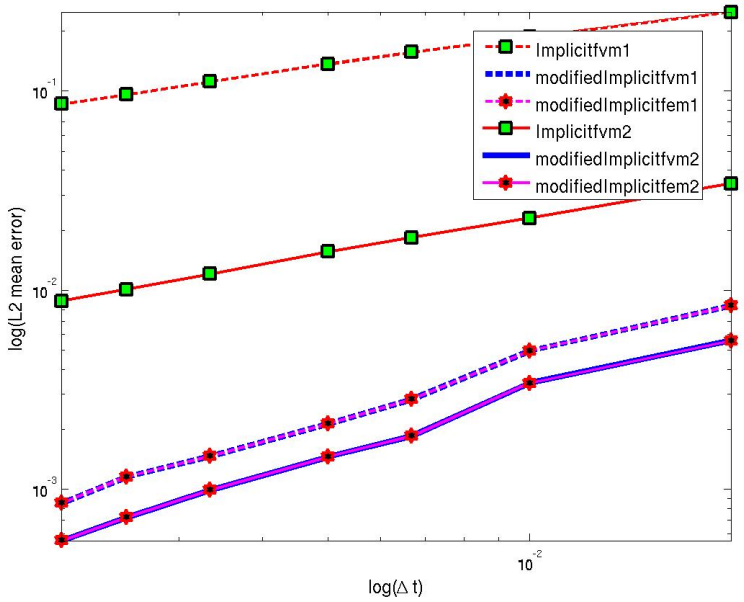
$$dX = \left(D\Delta X - \nabla \cdot (\mathbf{q}X) - \frac{X}{|X| + 1} \right) dt + dW, \tag{5.41}$$

with mixed Neumann-Dirichlet boundary conditions, we take $\Omega = [0, 1] \times [0, 1]$ and $D = 1/100$. The Dirichlet boundary condition is $X = 1$ at $x = 0$ and we use the homogeneous Neumann boundary conditions elsewhere. We consider a homogeneous media where the velocity is constant $\mathbf{q} = (1, 0)$. More realistic media, e.g. with highly fractured domains are considered in Chapter 6 and Chapter 7. In terms of equation (5.1) the nonlinear term F for physical values of X is given by $F(u) = -\nabla \cdot (\mathbf{q}u) - u/(|u| + 1)$, $u \in \mathbb{R}^+$ and clearly satisfies Assumption 5.3 (b). To easily deal with high Péclet flows we discretize in space using finite volumes. We can write the semi–discrete finite volume method as

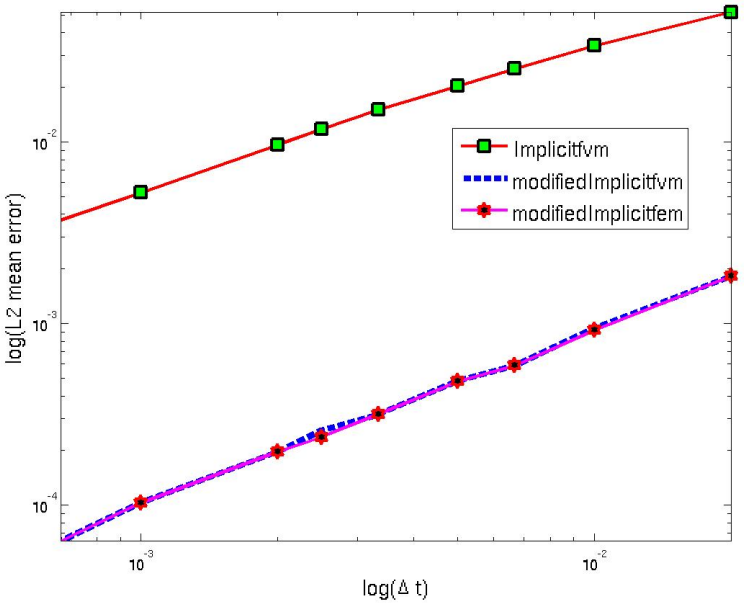
$$dX^h = (A_h X^h + P_h F(X^h) + b(X^h)) + P_h P_N dW, \tag{5.42}$$

where here A_h is the space discretization of $D\Delta$ using only homogeneous Neumann boundary conditions and $b(X^h)$ comes from the approximation of diffusion flux at the Dirichlet boundary condition size. Thus we can form the noise as in Section 5.4.1.

Figure 5.2 shows the convergence of the modified method with the noise in H^r, $r = 1, 2$ in space (denoted "ModifiedImplicitfvm$r$"). We observe that the temporal convergence order is close to 1/4. The predicted order 0.5 is not achieved probably due to the fact that the condition $F(X(t)) \in L_2(\mathbb{D}, \mathbb{H})$ is not satisfied. Figure 5.3(a) shows the mean of 20 realizations of the "true solution" (with the smallest time step $\Delta t = 1/7680$) for $r = 1$ while Figure 5.3(b) shows a sample of the "true solution".



(a)



(b)

Figure 5.1: Convergence in the root mean square L^2 norm at $T = 1$ as a function of Δt. (a) Shows convergence for noise both in H^r, $r = 1, 2$ (i.e. $O(t) \in L_2(\mathbb{D}, H^r(\Omega))$) for finite element and finite volume discretizations. We also show convergence of the standard semi–implicit scheme for the finite volume discretization. In (b) we show convergence for exponential correlation in the noise and the standard semi–implicit scheme is compared to the modified scheme. The initial solution is $X_0 = 0, \Gamma = 1$.

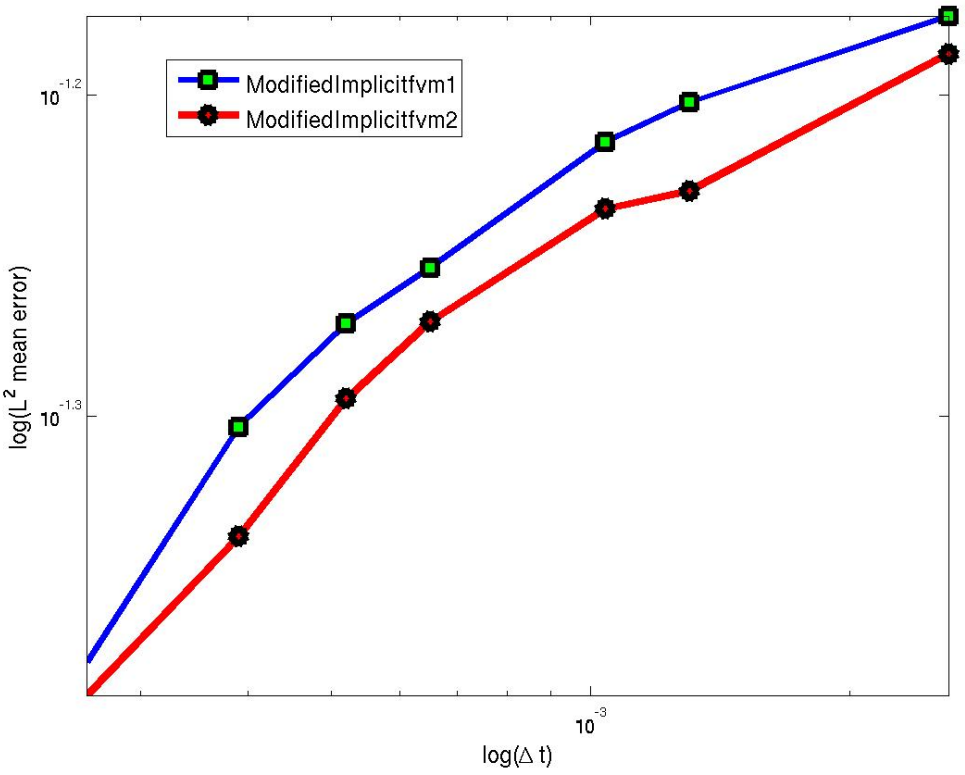
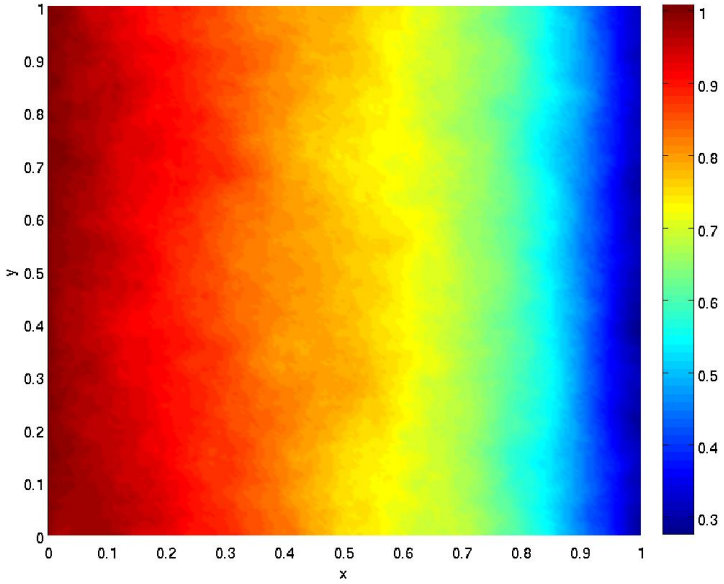
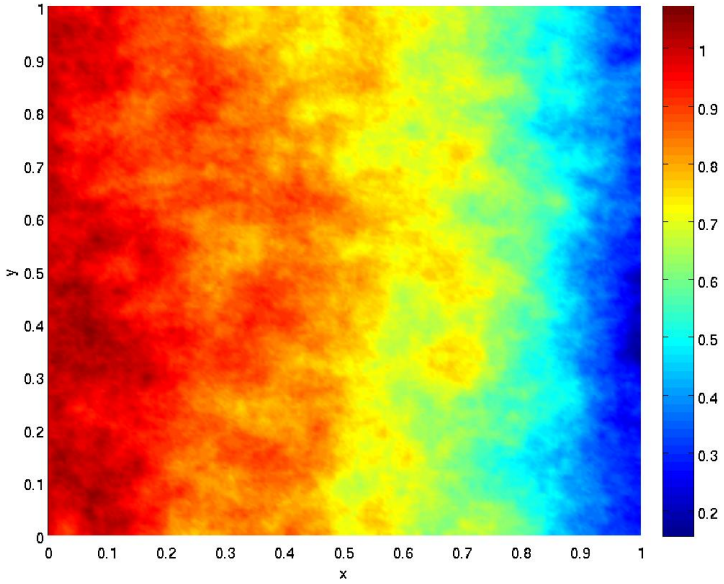


Figure 5.2: Convergence of the root mean square L^2 norm at $T = 1$ as a function of Δt with 20 realizations with $\Delta x = \Delta y = 1/200$, $X_0 = 0$, $\Gamma = 0.001$. The noise is white in time and in H^r in space, $r = 1, 2$ (i.e. $O(t) \in L_2(\mathbb{D}, H^r(\Omega))$). The temporal order of convergence in time is $1/4$

.



(a)



(b)

Figure 5.3: (a) Mean of the "true solution" for 20 realizations, $\Delta x = \Delta y = 1/200$, $X_0 = 0$, $\Gamma = 0.001$. The noise is white in time and in H^1 in space (i.e. $O(t) \in L_2(\mathbb{D}, H^1(\Omega))$). In (b) we show a sample of the "true solution" with $\Delta t = 1/7680$.

Chapter 6

Stochastic Exponential Integrators for a Finite Element Discretization of SPDEs with Additive Noise

In this chapter, we consider the numerical approximation of general semilinear parabolic stochastic partial differential equations (SPDEs) driven by additive space-time noise. In contrast to the standard time stepping methods which use basic increments of the noise and the approximation of the exponential operator (semigroup) by a rational fraction operator in the mild form, we introduce two new schemes, designed for finite element, finite volume or finite difference space discretization, similar to the schemes in [82, 89] for the spectral method and the scheme in Chapter 5 and [18] for the finite element method.

Our conditions on the noise and the nonlinear function are as in Chapter 5 and we consider the numerical examples from there as well as a more challenging example with stochastic flow in heterogeneous porous media. For the exponential integrators we rely on computing the exponential of a non-diagonal matrix. In our numerical results we use two different efficient techniques: the real fast Léja points and Krylov subspace techniques studied in Chapter 3. The results of this chapter are presented in our paper [19].

6.1 Introduction

As in Chapter 5, we consider the strong numerical approximation of Ito stochastic partial differential equations given in (5.1). Recent work by Jentzen and co-workers [82–84, 89] uses the Taylor expansion and linear functionals of the noise for Fourier–Galerkin discretizations of (4.11). In these schemes the diagonalization of the operator A through the discretization plays a key role. Using a linear functional of the noise overcomes the order barrier encountered using a standard increment of Wiener process [82]. In Chapter 5 and in [18] the use of linear functionals of the noise is extended to finite–element discretizations (where the operator does not diagonalize) with a semi-implicit Euler–Maruyama method. In contrast to the scheme presented in Chapter 5 and [18], here we consider two exponential based methods for time-stepping as in [81, 82, 89–91]. We prove a strong convergence result for two versions of the scheme with noise that is white in time and in H^1 and H^2 in space that shows that the exponential integrators are more accurate than the semi-implicit Euler-Maruyama method. Furthermore we have weaker restrictions on the regularity of the initial data and high accuracy for linear problems comparing to the scheme in Chapter 5 and in [18]. The cost of the extra accuracy though is that to implement these methods we need to compute the exponential functions of a non–diagonal matrix. Compared to the Fourier-Galerkin methods of [82–84, 89] we gain the flexibility of finite element (or finite volume method) to deal with complex boundary conditions and we can apply well developed techniques such as upwinding to deal with advection.

This chapter is organised as follows. In Section 6.2 we present the two numerical schemes based on the exponential integrators and our assumptions on (5.1). We present and comment on our convergence results. In Section 6.3 we present the proofs of our convergence theorems. We conclude in Section 6.4 by presenting some simulations and discuss implementation of these methods.

6.2 Numerical scheme and main results

We use the same notation and same functional spaces as in the previous chapter. Under the same technical assumptions as in Chapter 5 the unique mild solution of (5.1) is given

by

$$X(t) \;=\; S(t)X_0 + \int_0^t S(t-s)F(X(s))ds + O(t), \tag{6.1}$$

with the stochastic process O given by the stochastic convolution

$$O(t) = \int_0^t S(t-s)dW(s). \tag{6.2}$$

As in the previous chapter, we consider discretization of the spatial domain by a finite element triangulation. Recall that the semi-discretized version of (5.1) is to find the process $X^h(t) = X^h(.,t) \in V_h$ such that

$$dX^h = (A_h X^h + P_h F(X^h))dt + P_h P_N dW, \quad X^h(0) = P_h X_0. \tag{6.3}$$

The mild solution of (6.3) at time $t_m = m\Delta t$, $\Delta t > 0$ is given by

$$X^h(t_m) = S_h(t_m)P_h X_0 + \int_0^{t_m} S_h(t_m - s)P_h F(X^h(s))ds + \int_0^{t_m} S_h(t_m - s)P_h dW^N(s).$$

Given the mild solution at the time t_m, we can construct the corresponding solution at t_{m+1} as

$$\begin{aligned} X^h(t_{m+1}) \;&=\; S_h(\Delta t)X^h(t_m) + \int_0^{\Delta t} S_h(\Delta t - s)P_h F(X^h(s + t_m))ds \\ &\quad + \int_{t_m}^{t_{m+1}} S_h(t_{m+1} - s)P_h dW^N(s). \end{aligned} \tag{6.4}$$

For our first numerical scheme SETD1, we use the following approximations

$$F(X^h(t_m + s)) \;\approx\; F(X^h(t_m)) \qquad s \in [0,\ \Delta t],$$

and

$$\begin{aligned} \int_{t_m}^{t_{m+1}} S_h(t_{m+1} - s)P_h dW^N(s) \;&\approx\; P_h \int_{t_m}^{t_{m+1}} S_N(t_{m+1} - s)dW^N(s) \\ &=\; P_h P_N \int_{t_m}^{t_{m+1}} S(t_{m+1} - s)dW(s), \end{aligned}$$

where

$$A_N = P_N A \quad \text{and} \quad S_N(t) := e^{tA_N}.$$

Then we approximate X_m^h of $X(m\Delta t)$ by

$$\begin{aligned} X_{m+1}^h &= e^{\Delta t A_h} X_m^h + \Delta t \varphi_1(\Delta t A_h) P_h F(X_m^h) \\ &+ P_h \int_{t_m}^{t_{m+1}} e^{(t_{m+1}-s)A_N} dW^N(s) \end{aligned} \tag{6.5}$$

where

$$\varphi_1(\Delta t A_h) = (\Delta t\, A_h)^{-1} \left(e^{\Delta t A_h} - I\right) = \frac{1}{\Delta t} \int_0^{\Delta t} e^{(\Delta t - s)A_h} ds.$$

For efficiency to avoid computing two matrix exponentials, as in Chapter 3 we can rewrite the scheme (6.5) as

$$X_{m+1}^h = X_m^h + \Delta t \varphi_1(\Delta t A_h) \left(A_h X_m^h + P_h F(X_m^h)\right) + P_h \int_{t_m}^{t_{m+1}} e^{(t_{m+1}-s)A_N} dW^N(s).$$

We call this scheme SETD1.

Our second numerical method SETD0 is similar to the one in [81, 91]. It is based on approximating the deterministic integral in (6.4) at the left–hand endpoint of each partition and the stochastic integral as follows

$$\begin{aligned} \int_{t_m}^{t_{m+1}} S_h(t_{m+1}-s) P_h dW^N(s) &\approx P_h \int_{t_m}^{t_{m+1}} S_N(t_{m+1}-s) dW^N(s) \\ &= P_h P_N \int_{t_m}^{t_{m+1}} S(t_{m+1}-s) dW(s). \end{aligned}$$

With this we can define the SETD0 approximation Y_m^h of $X(m\Delta t)$ by

$$Y_{m+1}^h = \varphi_0(\Delta t A_h) \left(Y_m^h + \Delta t P_h F(Y_m^h)\right) + P_h \int_{t_m}^{t_{m+1}} e^{(t_{m+1}-s)A_N} dW^N(s) \tag{6.6}$$

where

$$\varphi_0(\Delta t A_h) = e^{\Delta t A_h}.$$

If we project the eigenfunctions of Q onto the eigenfunctions of the linear operator A then by a Fourier spectral method the process

$$\widehat{O}_k = \int_{t_k}^{t_{k+1}} e^{(t_{k+1}-s)A_N} dW^N(s)$$

is reduced to an Ornstein–Uhlenbeck process in each Fourier mode as in [82] and we therefore know the exact variance in each mode. We comment further on the implementation in Section 6.4.

For our convergence proofs, the assumptions that we make on the linear operator A, the nonlinear term F and the noise dW are the same as in Chapter 5. Note that for convenience of presentation we take A to be a second order operator. Similar results hold, however, for higher order operators.

6.2.1 Main results

Recall that as in Chapter 5, we let N be the number of terms of truncated noise, $\mathcal{I}_N = \{1, 2, ..., N\}^d$ and take $t_m = m\Delta t \in (0, T]$, where $T = M\Delta t$ for $m, M \in \mathbb{N}$. We take C to be a constant that may depend on T and other parameters but not on Δt, N or h. As in Chapter 5, we also assume that when initial data $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$ then $\mathbf{E}\|(-A)^\gamma X_0\|^4 < \infty$ with $0 \leq \gamma < 1$, which implies the regularity of the mild solution (5.1) (see Remark 5.6).

Our first result is a strong convergence result in L^2 when the non-linearity satisfies the Lipschitz condition of Assumption 5.3 (a) with scheme SETD1. This is, for example, the case of reaction–diffusion SPDEs.

Theorem 6.1 *Suppose that Assumptions 5.2, 5.3(a) and 5.5 (with $r = 1, 2$) are satisfied. Let $X(t_m)$ be the mild solution of equation (5.1) represented by (5.2) and X_m^h be the numerical approximation through scheme (6.5) (SETD1 scheme). Let $0 < \gamma < 1$ and set $\sigma = \min(2\theta, \gamma)$ and let $\theta \in (0, 1/2]$ be defined as in Assumption 5.5. If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $0 < \gamma < 1/2$ then*

$$\left(\mathbf{E}\|X(t_m) - X_m^h\|^2\right)^{1/2} \leq C\left(t_m^{-1/2+\gamma}h + \Delta t^\sigma + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-r/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $1/2 \leq \gamma < 1$ then

$$\left(\mathbf{E}\|X(t_m) - X_m^h\|^2\right)^{1/2} \leq C\left(h + \Delta t^\sigma + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-r/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, D(-A))$ and $F(X(t)) \in L_2(\mathbb{D}, \mathcal{D}((-A)^\alpha))$, $\forall t \in [0, T]$ with $\sup_{0\leq s\leq T}(\mathbf{E}\|F(X(s))\|_\alpha^2) < \infty$, $\alpha \in (0, 1/2)$ small enough then

$$\left(\mathbf{E}\|X(t_m) - X_m^h\|^2\right)^{1/2} \leq C\left(h^2 + \Delta t^{2\theta} + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-r/2}\right).$$

Our first result for scheme SETD0 is a strong convergence result in L^2 when the non-linearity satisfies the Lipschitz condition of Assumption 5.3 (a).

Theorem 6.2 *Suppose that Assumptions 5.2, 5.3(a) and 5.5 (with $r = 1, 2$) are satisfied. Let $X(t_m)$ be the mild solution of equation (5.1) represented by (5.2) and Y_m^h be the numerical approximation through scheme (6.6) (SETD0 scheme). Let $0 < \gamma < 1$ and set $\sigma = \min(2\theta, \gamma)$ and let $\theta \in (0, 1/2]$ be defined as in Assumption 5.5. If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $0 \leq \gamma < 1/2$ then*

$$\left(\mathbf{E}\|X(t_m) - Y_m^h\|^2\right)^{1/2} \leq C\left(t_m^{-1/2+\gamma}h + \Delta t^\sigma + \Delta t\,|\ln(\Delta t)| + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-r/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $1/2 \leq \gamma < 1$ then

$$\left(\mathbf{E}\|X(t_m) - Y_m^h\|^2\right)^{1/2} \leq C\left(h + \Delta t^\sigma + \Delta t\,|\ln(\Delta t)| + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-r/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ and $F(X(t)) \in L_2(\mathbb{D}, \mathcal{D}((-A)^\alpha))$, $\forall t \in [0, T]$ with $\sup_{0\leq s\leq T}(\mathbf{E}\|F(X(s))\|_\alpha^2) < \infty$, $\alpha \in (0, 1/2)$ small enough then

$$\left(\mathbf{E}\|X(t_m) - Y_m^h\|^2\right)^{1/2} \leq C\left(h^2 + \Delta t^{2\theta} + \Delta t\,|\ln(\Delta t)| + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-r/2}\right).$$

For convergence in the mean square $H^1(\Omega)$ norm where the non-linearity satisfies the Lipschitz condition from $L^2(\Omega)$ norm to $H^1(\Omega)$ (Assumption 5.3 (b)) we can state results for SETD1 and SETD0 together.

Theorem 6.3 *Suppose that Assumptions 5.2, 5.3(b), 5.5 (a) (with $r = 2$) are satisfied and $F(X(t)) \in L_2(\mathbb{D}, \mathbb{H})$, $\forall t \in [0, T]$ with $\mathbf{E}\left(\sup_{0\leq s\leq T}\|F(X(s))\|_1\right)^2 < \infty$. Let X be the solution mild of equation (5.1) represented by equation (5.2) and ζ_m^h be the numerical approximations through scheme (6.5) or (6.6) ($\zeta_m^h = X_m^h$ for scheme SETD1 and $\zeta_m^h = Y_m^h$ for scheme SETD0). Let $0 < \gamma < 1$. Then we have the following: If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{(1+\gamma)/2}))$ then*

$$(\mathbf{E}\|X(t_m) - \zeta_m^h\|_{H^1(\Omega)}^2)^{1/2} \leq C\left((t_m^{-1/2}\,h + \Delta t^{\gamma/2}) + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-1/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ then

$$(\mathbf{E}\|X(t_m) - \zeta_m^h\|_{H^1(\Omega)}^2)^{1/2} \leq C\left((h + \Delta t^{1/2-\epsilon}) + \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-1/2}\right),$$

for very small $\epsilon \in (0, 1/2)$.

We note that this theorem covers the case of advection-diffusion-reaction SPDEs, such as that arising in our example from porous media.

As in Chapter 5, we remark that if we denote by N_h the number of vertices in the finite element mesh then it is well known (see for example [79]) that if $N \geq N_h$ then

$$\left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-1} \leq Ch^2 \qquad \text{and} \qquad \left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-1/2} \leq Ch.$$

As a consequence the estimates in Theorem 6.1, Theorem 6.2 and Theorem 6.3 can be expressed as functions of h and Δt only, and it is the error from the finite element approximation that dominates. If $N \leq N_h$ then it is the error from the projection P_N of the noise onto a finite number of modes that dominates.

From Theorem 6.3 we also get an estimate in the root mean square $L^2(\Omega)$ norm in the case that the nonlinear function F satisfies Assumption 5.3 (b). We cannot do the proof directly in $L^2(\Omega)$ due to the Lipschitz condition in Assumption 5.3 (b). Simulations for Theorem 6.3 will be done in $L^2(\Omega)$ since the discrete $L^2(\Omega)$ norm is easier to estmate for the different types of boundary conditions.

Finally if we compare these theorems to those in [18] and Chapter 5 for a modified semi-implicit Euler-Maruyama method then we see that using the exponential based integrators we have weaker conditions on the initial data and in particular the scheme SETD1 has better convergence properties.

6.3 Proofs of main results

6.3.1 Preparatory result

We examine the deterministic linear problem. Find $u \in V$ such that such that

$$u' = Au \qquad \text{given} \quad u(0) = v \qquad\qquad t \in (0, T]. \tag{6.7}$$

The corresponding semi-discretization in space is : find $u_h \in V_h$ such that

$$u_h' = A_h u_h$$

where $u_h^0 = P_h v$. Define the operator

$$T_h(t) := S(t) - S_h(t)P_h = e^{tA} - e^{tA_h}P_h \tag{6.8}$$

so that $u(t) - u_h(t) = T_h(t)v$.

Lemma 6.4 *The following estimates hold on the semi-discrete approximation of (6.7)*

$$\|u(t) - u_h(t)\| = \|T_h(t)v\| \le Ch^r t^{-(r-\beta)/2}\|v\|_\beta \ \text{ if } \ v \in \mathcal{D}((-A)^{\beta/2}), \quad (6.9)$$

$$\|u(t) - u_h(t)\|_{H^1(\Omega)} = \|T_h(t)v\|_{H^1(\Omega)} \le Cht^{-1/2}\|v\|_1 \quad \text{if} \quad v \in \mathbb{H}, \quad (6.10)$$

$$\|u(t) - u_h(t)\|_{H^1(\Omega)} = \|T_h(t)v\|_{H^1(\Omega)} \le Ch\|v\|_2 \quad \text{if} \quad v \in \mathcal{D}(-A), \quad (6.11)$$

for $r = 1, 2, \ 0 \le \beta \le r$ *where* r *is linked to (5.22).*

Proof. Estimates (6.10)–(6.11) are the special case of the proof of Theorem 3.1 in [32] where the nonlinearity is taken to be zero. For our case

$$u(t) = S(t)v,$$

and we have the following estimates for $t \in (0, T]$

$$\|u(t)\|_{H^s(\Omega)} \le Ct^{-(s-1)/2}\|v\|_1 \quad \text{if} \quad v \in \mathbb{H} \quad s = 1, 2,$$

$$\|u(t)\|_{H^2(\Omega)} \le C\|v\|_2 \quad \text{if} \quad v \in \mathcal{D}(-A),$$

$$\|u_t(t)\|_{H^s(\Omega)} \le Ct^{-1-(s-1)/2}\|v\|_1 \quad \text{if} \quad v \in \mathbb{H} \quad s = 0, 1,$$

$$\|u_t(t)\|_{H^s(\Omega)} \le Ct^{-s/2}\|v\|_2 \quad \text{if} \quad v \in \mathcal{D}(-A) \quad s = 0, 1.$$

Using these in the proof of [32, Theorem 3.2] gives the result.

Let us prove the estimate (6.9). The proof for $r = 2$ and $\beta = 0$ can be found in [17]. Let R_h be the Riesz representation operator $R_h : V \to V_h$ defined in Chapter 5 by equation (5.21). Then

$$u_h(t) - u(t) = (u_h(t) - R_h u(t)) + (R_h u(t) - u(t)) \equiv \theta(t) + \rho(t). \quad (6.12)$$

It is well known [32] that $A_h R_h = P_h A$. Indeed for $v \in \mathcal{D}(A), \ \chi \in V_h$ we have

$$\begin{aligned} (P_h A v, \chi) &= (Av, \chi) && \text{(by definition of } P_h) \\ &= (AR_h v, \chi) && \text{(by definition of } R_h) \\ &= (A_h R_h v, \chi) && \text{(since } R_h v \in V_h) \end{aligned}$$

thus $A_h R_h = P_h A$. We therefore have the following equation in θ (see [32])

$$\theta_t = A_h \theta - P_h D_t \rho.$$

Hence

$$\theta(t) = S_h(t)\theta(0) - \int_0^t S_h(t-s) P_h D_s \rho ds.$$

Splitting the integral up into two intervals and integrating by parts over the first interval yields

$$\begin{aligned} \theta(t) &= S_h(t)\theta(0) + S_h(t)P_h\rho(0) - S_h(t/2)P_h\rho(t/2) + \int_0^{t/2} \left(D_s S_h(t-s)\right) P_h\rho(s)ds \\ &\quad - \int_{t/2}^t S_h(t-s)P_h D_s\rho(s)ds, \end{aligned}$$

with $D_s = \partial/\partial s$. Since $\theta(t) \in V_h$ we therefore have $P_h\theta(t) = \theta(t)$, then

$$\begin{aligned} \theta(t) &= S_h(t)P_h T_h(0)v - S_h(t/2)P_h\rho(t/2) + \int_0^{t/2} \left(D_s S_h(t-s)\right) P_h\rho(s)ds \\ &\quad - \int_{t/2}^t S_h(t-s)P_h D_s\rho(s)ds. \end{aligned}$$

Since

$$P_h T_h(0)v = P_h(v - P_h v) = 0,$$

we therefore have

$$\theta(t) = -S_h(t/2)P_h\rho(t/2) + \int_0^{t/2} D_s S_h(t-s)P_h\rho(s)ds - \int_{t/2}^t S_h(t-s)P_h D_s\rho(s)ds.$$

Using the fact that S_h and P_h are uniformly bounded independently of h with the smoothing property of S_h in Proposition 2.6 yields

$$\|\theta(t)\| \leq C\left(\|\rho(t/2)\| + \int_0^{t/2} (t-s)^{-1}\|\rho(s)\|ds + \int_{t/2}^t \|D_s\rho(s)\|ds\right).$$

Using (5.22) with the smoothing property of $S(t)$ in Proposition 2.6 yields

$$\begin{cases} \|\rho(t)\| \leq Ch^r\|u\|_r \leq Ch^r t^{-(r-\beta)/2}\|v\|_\beta \\ \\ \|D_s\rho(t)\| \leq Ch^r\|D_s u\|_r \leq Ch^r t^{-1-(r-\beta)/2}\|v\|_\beta, \quad r \in \{1,2\}, \ \beta \leq r, \quad v \in \mathcal{D}((-A)^{\beta/2}). \end{cases}$$

Then

$$\|\theta(t)\| \leq Ch^r t^{-(r-\beta)/2}\|v\|_\beta + Ch^r\|v\|_\beta \left(\int_0^{t/2} (t-s)^{-1} s^{-(r-\beta)/2} ds + \int_{t/2}^{t} s^{-1-(r-\beta)/2} ds \right).$$

Since

$$\int_0^{t/2} (t-s)^{-1} s^{-(r-\beta)/2} ds + \int_{t/2}^{t} s^{-1-(r-\beta)/2} ds \leq Ct^{-(r-\beta)/2},$$

we therefore have

$$\|T_h(t)v\| \leq \|\theta(t)\| + \|\rho(t)\| \leq Ch^r\ t^{-(r-\beta)/2}\|v\|_\beta.$$

■

6.3.2 Proof of Theorem 6.1

The proof follows the same basic steps as in Chapter 5 and [18], however here the discrete semigroup is an exponential. As a consequence the estimates are different and the proof here is simpler with fewer terms to estimate. Set

$$\begin{aligned} X(t_m) &= S(t_m)X_0 + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S(t_m - s)F(X(s))ds + O(t_m) \\ &= \overline{X}(t_m) + O(t_m). \end{aligned}$$

Recall that by construction

$$\begin{aligned} X_m^h &= e^{\Delta t A_h} X_{m-1}^h + \int_0^{\Delta t} e^{(\Delta t - s)A_h} P_h F(X_{m-1}^h) ds + P_h \int_{t_{m-1}}^{t_m} e^{(t_m - s)A_N} dW^N(s) \\ &= S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1} \left(\int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h F(X_k^h) ds + P_h \int_{t_k}^{t_{k+1}} S_N(t_m - s) dW^N(s) \right) \\ &= S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1} \left(\int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h F(X_k^h) ds \right) + P_h P_N O(t_m) \\ &= Z_m^h + P_h P_N O(t_m), \end{aligned}$$

where

$$\begin{aligned} Z_m^h &= S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1} \left(\int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h F(X_k^h) ds \right) \\ &= S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1} \left(\int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h F(Z_k^h + P_h P_N O(t_k)) ds \right). \end{aligned}$$

We now estimate $\left(\mathbf{E}\|X(t_m) - X_m^h\|^2\right)^{1/2}$. We obviously have

$$\begin{aligned} X(t_m) - X_m^h &= \overline{X}(t_m) + O(t_m) - X_m^h \\ &= \overline{X}(t_m) + O(t_m) - \left(Z_m^h + P_h P_N O(t_m)\right) \\ &= \left(\overline{X}(t_m) - Z_m^h\right) + (P_N(O(t_m)) - P_h P_N(O(t_m))) + (O(t_m) - P_N(O(t_m))) \\ &= I + II + III. \end{aligned} \tag{6.13}$$

Then $\left(\mathbf{E}\|X(t_m) - X_m^h\|^2\right)^{1/2} \leq (\mathbf{E}\|I\|^2)^{1/2} + (\mathbf{E}\|II\|^2)^{1/2} + (\mathbf{E}\|III\|^2)^{1/2}$ and we estimate each term. Since the first term will require the most work we first estimate the other two.

Let us estimate $(\mathbf{E}\|II\|^2)^{1/2}$. Using the property (5.23) of the projection P_h, the equivalence $\|.\|_{H^r(\Omega)} \equiv \|(-A)^{r/2}.\|$ in $\mathcal{D}((-A)^{r/2})$, the Ito isometry and the fact that the semigroup is a bounded operator yields

$$\begin{aligned} \mathbf{E}\|II\|^2 &\leq Ch^{2r}\mathbf{E}\|(-A)^{r/2}\int_0^{t_m} S(t_m - s)dW(s)\|^2 \\ &\leq Ch^{2r}\int_0^{t_m} \|(-A)^{r/2}S(t_m - s)\|_{L_2^0}^2 ds \\ &\leq Ch^{2r}\int_0^{T} \|(-A)^{r/2}Q^{1/2}\|_{HS}^2 ds. \end{aligned}$$

Thus, since the process $O(t) \in L_2(\mathbb{D}, H^r(\Omega))$ we have $(\mathbf{E}\|II\|^2)^{1/2} \leq Ch^r$.

For the third term III

$$\mathbf{E}\|III\|^2 = \mathbf{E}\|(\mathrm{I} - P_N)O(t_m)\|^2 = \mathbf{E}\|(\mathrm{I} - P_N)(-A)^{-r/2}(-A)^{r/2}O(t_m)\|^2,$$

and so

$$\mathbf{E}\|III\|^2 \leq \|(\mathrm{I} - P_N)(-A)^{-r/2}\|_{L(L^2(\Omega))}^2 \mathbf{E}\|(-A)^{r/2}O(t_m)\|^2 \leq C\left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N} \lambda_j\right)^{-r}.$$

We now turn our attention to the first term $\mathbf{E}\|I\|^2$. Using the definition of T_h from (6.8)

the first term I can be expanded

$$\begin{aligned} I &= T_h X_0 + \sum_{k=0}^{m-1} \int_{t_k t}^{t_{k+1}} S(t_m - s) F(X(s)) - S_h(t_m - s) P_h F(Z_k^h + P_h P_N O(t_k)) ds \\ &= T_h X_0 + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_h(t_m - s) P_h (F(X(t_k)) - F(Z_k^h + P_h P_N O(t_k)))) ds \\ &\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_h(t_m - s) P_h (F(X(s)) - F(X(t_k))) ds \\ &\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (S(t_m - s) - S_h(t_m - s) P_h) F(X(s)) ds \\ &= I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{6.14}$$

Then

$$\left(\mathbf{E}\|I\|^2\right)^{1/2} \leq \left(\mathbf{E}\|I_1\|^2\right)^{1/2} + \left(\mathbf{E}\|I_2\|^2\right)^{1/2} + \left(\mathbf{E}\|I_3\|^2\right)^{1/2} + \left(\mathbf{E}\|I_4\|^2\right)^{1/2}.$$

For I_1, from (6.9) of Lemma 6.4 we have:

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma)),\ 0 \leq \gamma < 1/2$

$$(\mathbf{E}\|I_1\|^2)^{1/2} \leq C t_m^{-1/2+\gamma} h \left(\mathbf{E}\|X_0\|_{2\gamma}^2\right)^{1/2}.$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma)),\ 1/2 \leq \gamma < 1$

$$(\mathbf{E}\|I_1\|^2)^{1/2} \leq C h \left(\mathbf{E}\|X_0\|_1^2\right)^{1/2}.$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$,

$$(\mathbf{E}\|I_1\|^2)^{1/2} \leq C h^2 \left(\mathbf{E}\|X_0\|_2^2\right)^{1/2}.$$

If F satisfies Assumption 5.3 (a), then using the Lipschitz condition, triangle inequality as well as that $S_h(t)$ and P_h are bounded operators, we have

$$\begin{aligned} \left(\mathbf{E}\|I_2\|^2\right)^{1/2} &\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|F(X(t_k)) - F((Z_k^h + P_h P_N O(t_k))\|^2\right)^{1/2} ds \\ &\leq C \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2\right)^{1/2} ds. \end{aligned}$$

As I_3 needs more work let us estimate I_4 first. Using the fact P_h, S, S_h are bounded with (6.9) of Lemma 6.4 yields

$$\begin{aligned}\left(\mathbf{E}\|I_4\|^2\right)^{1/2} &\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}\left(\mathbf{E}\|T_h(t_m-s)F(X(s))\|^2\right)^{1/2}ds \\ &\leq Ch\sup_{0\leq s\leq T}\left(\mathbf{E}\|F(X(s))\|^2\right)^{1/2}\left(\int_0^{t_m}(t_m-s)^{-1/2}\,ds\right) \\ &\leq Ch.\end{aligned}$$

If $F(X(t)) \in L_2(\mathbb{D}, \mathcal{D}((-A)^\alpha))$, with $\sup_{0\leq s\leq T}(\mathbf{E}\|F(X(s))\|_\alpha^2) < \infty$, $\alpha \in (0, 1/2)$ small enough, we also have

$$\begin{aligned}\left(\mathbf{E}\|I_4\|^2\right)^{1/2} &\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}\left(\mathbf{E}\|T_h(t_m-s)F(X(s))\|^2\right)^{1/2}ds \\ &\leq Ch^2\sup_{0\leq s\leq T}\left(\mathbf{E}\|F(X(s))\|_\alpha^2\right)^{1/2}\left(\int_0^{t_m}(t_m-s)^{-1+\alpha}\,ds\right) \\ &\leq Ch^2.\end{aligned}$$

Let us estimate $(\mathbf{E}\|I_3\|^2)^{1/2}$. We add in and subtract out $O(s)$ and $O(t_k)$ yields

$$\begin{aligned}&\left(\mathbf{E}\|I_3\|^2\right)^{1/2} \\ &= \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}S_h(t_m-s)P_h\left(F(X(s))-F(X(t_k)+O(s)-O(t_k))\right)ds\|^2\right)^{1/2} \\ &\quad+\left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}S_h(t_m-s)P_h\left(F(X(t_k)+O(s)-O(t_k))-F(X(t_k))\right)ds\|^2\right)^{1/2} \\ &:= \left(\mathbf{E}\|I_3^1\|^2\right)^{1/2}+\mathbf{E}\left(\|I_3^2\|^2\right)^{1/2}.\end{aligned}$$

Applying the Lipschitz condition in Assumption 5.3(a), using the fact that the semigroup is bounded and according to Lemma 5.10, for $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $\quad 0 \leq \gamma \leq 1$ we therefore have

$$\begin{aligned}\left(\mathbf{E}\|I_3^1\|^2\right)^{1/2} &\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}\left(\mathbf{E}\|(X(s)-O(s))-(X(t_k)-O(t_k))\|^2\right)^{1/2}ds \\ &\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(s-t_k)^\gamma ds \quad \leq \quad C\Delta t^\gamma.\end{aligned}$$

Let us now estimate $\mathbf{E}\left(\|I_3^2\|^2\right)^{1/2}$. The analysis below follows the same steps as in Chapter 5 although the approximating semigroup S_h is different here. Applying a Taylor expansion to F gives

$$\mathbf{E}\left(\|I_3^2\|^2\right)^{1/2} \leq I_3^{21} + I_3^{22} + I_3^{23},$$

with

$$\begin{aligned}
I_3^{21} &= \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_h(t_m-s)P_hF'(X(t_k))(O(s)-S(s-t_k)O(t_k))ds\|^2\right)^{1/2} \\
I_3^{22} &= \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_h(t_m-s)P_hF'(X(t_k))(S(s-t_k)O(t_k)-O(t_k))ds\|^2\right)^{1/2} \\
I_3^{23} &= \left(\mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_h(t_m-s)\int_0^1 G(1-r)drds\|^2\right)^{1/2}, \\
G &:= P_hF''(X(t_k)) + r(O(s)-O(t_k))(O(s)-O(t_k),O(s)-O(t_k)).
\end{aligned}$$

Using the fact that $O(t_2)-S(t_2-t_1)O(t_1)$, $0\leq t_1<t_2\leq T$ is independent of $\mathcal{F}_{t_1}$, one can show, as in [89], that

$$\left(I_3^{21}\right)^2 = \sum_{k=0}^{m-1}\mathbf{E}\|\int_{t_k}^{t_{k+1}} S_h(t_m-s)P_hF'(X(t_k))(O(s)-S(s-t_k)O(t_k))ds\|^2.$$

Therefore as S_h is bounded we have

$$\begin{aligned}
&I_3^{21} \\
&\leq \left(\sum_{k=0}^{m-1}\left(\int_{t_k}^{t_{k+1}}\left(\mathbf{E}\|S_h(t_m-s)P_hF'(X(t_k))(O(s)-S(s-t_k)O(t_k))\|^2\right)^{1/2}ds\right)^2\right)^{1/2} \\
&\leq C\left(\sum_{k=0}^{m-1}\left(\int_{t_k}^{t_{k+1}}\left(\mathbf{E}\|P_hF'(X(t_k))(O(s)-S(s-t_k)O(t_k))\|^2\right)^{1/2}ds\right)^2\right)^{1/2}.
\end{aligned}$$

Hölder's inequality with Assumption 5.5, Assumption 5.3(a) and Proposition 2.6 yields

$$
\begin{aligned}
& I_3^{21} \\
&\leq C\Delta t^{1/2}\left(\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \mathbf{E}\|P_hF'(X(t_k))(O(s)-S(s-t_k)O(t_k))\|^2 ds\right)^{1/2} \\
&\leq C\Delta t^{1/2}\left(\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \mathbf{E}\|(O(s)-S(s-t_k)O(t_k))\|^2 ds\right)^{1/2} \\
&\leq C\Delta t^{1/2}\left(\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \left(\left(\mathbf{E}\|O(s)-O(t_k)\|^2\right)^{1/2}+\left(\mathbf{E}\|(S(s-t_k)-\mathbf{I})O(t_k))\|^2\right)^{1/2}\right)^2 ds\right)^{1/2} \\
&\leq C\Delta t^{1/2}\left(\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \left((s-t_k)^{\theta}+(s-t_k)^{r/2}\left(\mathbf{E}\|O(t_k)\|_r^2\right)^{1/2}\right)^2 ds\right)^{1/2} \\
&\leq C\Delta t^{1/2+\theta}.
\end{aligned}
$$

Let us estimate I_3^{22}.

$$
\begin{aligned}
& I_3^{22} \\
&\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|S_h(t_m-s)P_h(-A)^{1/2}(-A)^{-1/2}F'(X(t_k))(S(s-t_k)-\mathbf{I})O(t_k))\|^2\right)^{1/2} ds \\
&\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \|S_h(t_m-s)P_h(-A)^{1/2}\|_{L(L^2(\Omega))} \\
&\quad \times \left(\mathbf{E}\|(-A)^{-1/2}F'(X(t_k))(S(s-t_k)-\mathbf{I})O(t_k))\|^2\right)^{1/2} ds.
\end{aligned}
$$

Since $P_h(-A)^{1/2}=(-A_h)^{1/2}$ and S_h satisfies the smoothing properties analogous to $S(t)$ independently of h (see for example [32]), and in particular

$$
\|S_h(t_m)(-A_h)^{1/2}\|_{L(L^2(\Omega))}=\|(-A_h)^{1/2}S_h(t_m)\|_{L(L^2(\Omega))}\leq Ct_m^{-1/2}, \quad t_m=m\Delta t>0,
$$

we therefore have

$$
\begin{aligned}
I_3^{22} &\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m-s)^{-1/2} \\
&\quad \times \left(\mathbf{E}\|(-A)^{-1/2}F'(X(t_k))((S(s-t_k)-\mathbf{I})O(t_k))\|^2\right)^{1/2} ds.
\end{aligned}
$$

As in the similar estimate of I_3^{22} in Chapter 5, the identification of H to its dual yields

$$
\begin{aligned}
I_3^{22} &\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-t_{k+1})^{-1/2} \\
&\quad\times\left(\mathbf{E}\left(\sup_{\|v\|\leq 1}|\langle F'(X(t_k))^*(-A)^{-1/2}v,(S(s-t_k)-\mathbf{I})O(t_k)\rangle|\right)^2\right)^{1/2}ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-t_{k+1})^{-1/2} \\
&\quad\times\left(\mathbf{E}\left(\sup_{\|v\|\leq 1}\|F'(X(t_k))^*(-A)^{-1/2}v\|_1\|(S(s-t_k)-\mathbf{I})O(t_k))\|_{-1}\right)^2\right)^{1/2}ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-t_{k+1})^{-1/2} \\
&\quad\times\left(\mathbf{E}\left(1+\|X(t_k)\|_1\right)^2\|\left(S(s-t_k)-\mathbf{I}\right)O(t_k)\|_{-1}\right)^2\right)^{1/2}ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(t_m-t_{k+1})^{-1/2} \\
&\quad\times\left(\mathbf{E}\left(1+\|X(t_k)\|_1\right)^4\right)^{1/4}\left(\mathbf{E}\left(\|S(s-t_k)-\mathbf{I})O(t_k)\|_{-1}\right)^4\right)^{1/4}ds \\
&\leq C\sum_{k=0}^{m-1}(t_m-t_{k+1})^{-1/2} \\
&\quad\times\left(1+\left(\mathbf{E}\|X(t_k)\|_1^4\right)^{1/4}\right)\int_{t_k}^{t_{k+1}}\left(\mathbf{E}\left(\|S(s-t_k)-\mathbf{I})O(t_k)\|_{-1}\right)^4\right)^{1/4}ds \\
&\leq C\sum_{k=0}^{m-1}(t_m-t_{k+1})^{-1/2} \\
&\quad\times\int_{t_k}^{t_{k+1}}\|(-A)^{-(r/2+1/2)}\left(S(s-t_k)-\mathbf{I}\right)\|_{L(L^2(\Omega))}\left(\mathbf{E}\|O(t_k)\|_r^4\right)^{1/4}ds \\
&\leq C\sum_{k=0}^{m-1}(t_m-t_{k+1})^{-1/2}\int_{t_k}^{t_{k+1}}\|(-A)^{1/2-r/2}(-A)^{-1}(S(s-t_k)-\mathbf{I})\|_{L(L^2(\Omega))}ds.
\end{aligned}
$$

Using Proposition 2.6 and the fact that $(-A)^{1/2-r/2}$ is bounded as $r = 1, 2$ yields

$$\begin{aligned} I_3^{22} &\leq C\sum_{k=0}^{m-1} (t_m - t_{k+1})^{-1/2} \int_{t_k}^{t_{k+1}} (s - t_k)\, ds \\ &= C\Delta t^{3/2} \sum_{k=0}^{m-1} (m-k-1)^{-1/2}. \end{aligned}$$

We can bound the sum above by $2M^{1/2}$, therefore we have

$$I_3^{21} + I_3^{22} \leq C(\Delta t + \Delta t^{1/2+\theta}) \leq C(\Delta t^{2\theta}).$$

Let us estimate I_3^{23}. Using the fact that S_h is bounded with Assumption 5.3 and Assumption 5.5 yields (with $G = P_h F''(X(t_k) + r(O(s) - O(t_k)))(O(s) - O(t_k), O(s) - O(t_k)))$

$$\begin{aligned} I_3^{23} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \|S_h(t_m - s)\|_{L(L^2(\Omega))} \int_0^1 \left(\mathbf{E}\|G\|^2\right)^{1/2} dr ds \\ &\leq C\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \int_0^1 \left(\mathbf{E}\|O(s) - O(t_k)\|_{\mathcal{V}}^4\right)^{1/2} dr ds \\ &\leq C\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\left(\mathbf{E}\|O(s) - O(t_k)\|_{\mathcal{V}}^4\right)^{1/4}\right)^2 ds \\ &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (s - t_k)^{2\theta}\, ds \leq C(\Delta t)^{2\theta}. \end{aligned}$$

Combining $I_3^{21} + I_3^{22}$ and I_3^{23} yields the following estimate

$$\mathbf{E}\left(\|I_3\|^2\right)^{1/2} \leq C(\Delta t^{\sigma}) \leq C(\Delta t^{\sigma}).$$

Combining the previous estimates for the term I yields :

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\gamma})),\ 0 < \gamma < 1/2$

$$\left(\mathbf{E}\|I\|^2\right)^{1/2} \leq C\left(t_m^{-1/2+\gamma} h + \Delta t^{\sigma} + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2\right)^{1/2} ds\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\gamma})),\ 1/2 \leq \gamma < 1$

$$\left(\mathbf{E}\|I\|^2\right)^{1/2} \leq C\left(h + \Delta t^{\sigma} + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2\right)^{1/2} ds\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ and $F(X(t)) \in L_2(\mathbb{D}, \mathcal{D}((-A)^\alpha))$, with $\sup_{0\le s\le T}(\mathbf{E}\|F(X(s))\|_\alpha^2) < \infty$, $\alpha \in (0, 1/2)$ small enough,

$$\left(\mathbf{E}\|I\|^2\right)^{1/2} \;\le\; C\left(h^2 + \Delta t^{2\theta} + \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2\right)^{1/2} ds\right).$$

Finally we combine all our estimates on I, II and III to get $(\mathbf{E}\|I\|^2)^{1/2}$, $(\mathbf{E}\|II\|^2)^{1/2}$ and $(\mathbf{E}\|III\|^2)^{1/2}$ and use the discrete Gronwall lemma to complete the proof.

6.3.3 Proof of Theorem 6.3 for SETD1 scheme

We now prove convergence in $H^1(\Omega)$ and estimate $\left(\mathbf{E}\|X(t_m) - X_m^h\|_{H^1(\Omega)}^2\right)^{1/2}$. For the proof we follow the same steps as in previous section for Theorem 6.1 and we now estimate (6.13) in the H^1 norm.

Let estimate $(\mathbf{E}\|II\|_{H^1(\Omega)}^2)^{1/2}$. As in the proof of Theorem 6.1 in Section 6.3.2, using the regularity of the noise $O(t) \in L_2(\mathbb{D}, \mathcal{D}(-A))$, $\forall t \in [0, T]$ and the property (5.23) of the projection P_h yields

$$\begin{aligned}
\mathbf{E}\|II\|_{H^1(\Omega)}^2 &= \mathbf{E}\|P_h P_N(O(t_m)) - P_N(O(t_m))\|_{H^1(\Omega)}^2 \\
&\le Ch^2\mathbf{E}\|(P_N(O(t_m))\|_{H^2(\Omega)} \\
&\le Ch^2\mathbf{E}\|O(t_m)\|_{H^2(\Omega)} \\
&\le Ch^2\mathbf{E}\|(-A)\int_0^{t_m} S(t_m - s)dW(s)\|^2
\end{aligned}$$

$$\begin{aligned}
\mathbf{E}\|II\|_{H^1(\Omega)}^2 &\le Ch^2\int_0^{t_m} \|(-A)S(t_m - s)\|_{L_2^0}^2 ds \\
&\le Ch^2\int_0^{T} \|(-A)Q^{1/2}\|_{HS}^2 ds \\
&\le Ch^2,
\end{aligned}$$

thus $(\mathbf{E}\|II\|_{H^1(\Omega)}^2)^{1/2} \le Ch$.

Using the regularity of the noise again and the equivalency $\|.\|_{H^1(\Omega)} \equiv \|(-A)^{1/2}.\|$, we

also have

$$\begin{aligned}
\mathbf{E}\|III\|^2_{H^1(\Omega)} &= \mathbf{E}\|(\mathrm{I}-P_N)O(t_m)\|^2_{H^1(\Omega)} \\
&= \mathbf{E}\|(\mathrm{I}-P_N)(-A)^{-1}(-A)^1O(t_m)\|^2_{H^1(\Omega)} \\
&= \mathbf{E}\|(-A)^{1/2}(\mathrm{I}-P_N)(-A)^{-1}(-A)^1O(t_m)\|^2_{H^1(\Omega)} \\
&\leq \|(-A)^{1/2}(\mathrm{I}-P_N)(-A)^{-1}\|^2_{L(L^2(\Omega))}\mathbf{E}\|(-A)O(t_m)\|^2 \\
&\leq \|(-A)^{1/2}(\mathrm{I}-P_N)(-A)^{-1}\|^2_{L(L^2(\Omega))}\mathbf{E}\|(-A)O(t_m)\|^2 \\
&\leq C\left(\inf_{j\in\mathbb{N}^d\setminus\mathcal{I}_N}\lambda_j\right)^{-1}.
\end{aligned}$$

We now estimate the term I from (6.13) in the $H^1(\Omega)$ norm noting that from (6.14) we have $I = I_1 + I_2 + I_3 + I_4$. Estimates on I_1 follow immediately from equations (6.10) and (6.11) of Lemma 6.4, and then for I_1, if $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{(\gamma+1)/2}))$, equation (6.10) of Lemma 6.4 gives

$$(\mathbf{E}\|I_1\|^2_{H^1(\Omega)})^{1/2} \leq Ct_m^{-1/2}h\left(\mathbf{E}\|X_0\|_1^2\right)^{1/2}$$

and if $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$,

$$(\mathbf{E}\|I_1\|^2_{H^1(\Omega)})^{1/2} \leq Ch\left(\mathbf{E}\|X_0\|_2^2\right)^{1/2}.$$

If F satisfies Assumption 5.3 (b), then using the Lipschitz condition, the triangle inequality, the fact that P_h is an bounded operator and S_h satisfies the smoothing property analogous to $S(t)$ independently of h [32], i.e.

$$\|S_h(t)v\|^2_{H^1(\Omega)} \leq Ct^{-1/2}\|v\| \qquad v \in V_h \quad t > 0,$$

we have

$$
\begin{aligned}
&(\mathbf{E}\|I_2\|^2_{H^1(\Omega)})^{1/2} \\
&\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|S_h(t_m-s)P_h\left(F(X(t_k))-F(Z_k^h+P_hP_NO(t_k))\right)\|^2_{H^1(\Omega)}\right)^{1/2} ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m-s)^{-1/2}\left(\mathbf{E}\|F(X(t_k))-F(Z_k^h+P_hP_NO(t_k))\|^2\right)^{1/2} ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m-s)^{-1/2}\left(\mathbf{E}\|X(t_k)-X_k^h\|^2_{H^1(\Omega)}\right)^{1/2} ds.
\end{aligned}
$$

Once again using the Lipschitz condition, triangle inequality and smoothing property of S_h, but with Lemma 5.10 gives

$$
\begin{aligned}
&(\mathbf{E}\|I_3\|^2_{H^1(\Omega)})^{1/2} \\
&\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (\mathbf{E}\|S_h(t_m-s)P_h(F(X(s))-F(X(t_k))\|^2_{H^1(\Omega)})^{1/2})ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m-s)^{-1/2}(\mathbf{E}\|F(X(s))-F(X(t_k))\|)^{1/2}ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m-s)^{-1/2}(\mathbf{E}\|X(s)-X(t_k)\|^2_{H^1(\Omega)})^{1/2}ds \\
&\leq C\left(\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m-s)^{-1/2}(s-t_k)^{\gamma/2}ds\right) \\
&\quad\times\left(\mathbf{E}\|X_0\|^2_{\gamma+1}+\left(\mathbf{E}\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2+1\right)^{1/2} \\
&\leq C\left(\Delta t^{\gamma/2}\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (t_m-s)^{-1/2}\right) \\
&\quad\times\left(\mathbf{E}\|X_0\|^2_{\gamma+1}+\left(\mathbf{E}\sup_{0\leq s\leq T}\|F(X(s))\|_{H^1(\Omega)}\right)^2+1\right)^{1/2}.
\end{aligned}
$$

As in the previous theorem, we use the fact that

$$\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} ds \leq 2\sqrt{T}.$$

Then if $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{(\gamma+1)/2}))$ we have finally found

$$(\mathbf{E}\|I_3\|^2_{H^1(\Omega)})^{1/2} \leq C(\Delta t)^{\gamma/2} \left(\mathbf{E}\|X_0\|^2_{\gamma+1} + \left(\mathbf{E} \sup_{0\leq s\leq T} \|F(X(s))\|_{H^1(\Omega)} \right)^2 + 1 \right)^{1/2}.$$

In the same way, if $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ we obviously have $(\mathbf{E}\|I_3\|^2_{H^1(\Omega)})^{1/2} \leq C(\Delta t)^{1/2-\epsilon}$ by taking $\gamma = 1 - \epsilon$ in Lemma 5.10, $\epsilon > 0$ small enough.

If $F(X(t)) \in L_2(\mathbb{D}, \mathbb{H})$ then by (6.10) of Lemma 6.4 we find

$$\begin{aligned}
(\mathbf{E}\|I_4\|^2_{H^1(\Omega)})^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|T_h(t_m - s)F(X(s))\|^2_{H^1(\Omega)} \right)^{1/2} ds \\
&\leq Ch \left(\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} ds \right) \left(\sup_{0\leq s\leq T} \mathbf{E}\|F(X(s))\|^2_{H^1(\Omega)} \right)^{1/2} \\
&\leq Ch.
\end{aligned}$$

Combining our estimates, for $F(X(t)) \in L_2(\mathbb{D}, \mathbb{H})$ we have that: If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{(\gamma+1)/2}))$ then

$$\begin{aligned}
(\mathbf{E}\|I\|^2_{H^1(\Omega)})^{1/2} &\leq C \left(t_m^{-1/2} h + \Delta t^{\gamma/2} \right. \\
&+ \left. \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2_{H^1(\Omega)} \right)^{1/2} ds \right).
\end{aligned}$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ then

$$\begin{aligned}
(\mathbf{E}\|I\|^2_{H^1(\Omega)})^{1/2} &\leq C \left(h + \Delta t^{(\frac{1}{2}-\epsilon)} \right. \\
&+ \left. \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (t_m - s)^{-1/2} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2_{H^1(\Omega)} \right)^{1/2} ds \right).
\end{aligned}$$

where $C > 0$ depending of the T, the initial solution X_0, the mild solution X, the nonlinear function F.

Combining our estimates $\left(\mathbf{E}\|I\|^2_{H^1(\Omega)} \right)^{1/2}$, $\left(\mathbf{E}\|II\|^2_{H^1(\Omega)} \right)^{1/2}$ and $\left(\mathbf{E}\|III\|^2_{H^1(\Omega)} \right)^{1/2}$ and using the discrete Gronwall lemma concludes the proof.

6.3.4 Proofs of Theorem 6.2 and Theorem 6.3 for SETD0 scheme

Recall that

$$\begin{aligned}
Y_m^h &= e^{\Delta t A_h}\left(Y_{m-1}^h + \Delta t P_h F(Y_{m-1}^h)\right) + P_h \int_{t_{m-1}}^{t_m} e^{(t_m-s)A_N} dW^N(s) \\
&= S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1}\left(\int_{t_k}^{t_{k+1}} S_h(t_m - t_k)P_h F(Y_k^h)ds \right. \\
&\quad \left. + P_h \int_{t_k}^{t_{k+1}} S_N(t_m - s)dW^N(s)\right) \\
&= S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1}\left(\int_{t_k}^{t_{k+1}} S_h(t_m - t_k)P_h F(Y_k^h)ds\right) + P_h P_N O(t_m) \\
&= Z_m^h + P_h P_N O(t_m).
\end{aligned}$$

As in the Theorem 6.1 we obviously have

$$\begin{aligned}
&X(t_m) - Y_m^h \\
&= \overline{X}_{t_m} + O(t_m) - Y_m^h \\
&= \overline{X}(t_m) + O(t_m) - \left(Z_m^h + P_h P_N O(t_m)\right) \\
&= \left(\overline{X}(t_m) - Z_m^h\right) + (P_N(O(t_m)) - P_h P_N(O(t_m))) + (O(t_m) - P_N(O(t_m))) \\
&= I + II + III.
\end{aligned} \tag{6.15}$$

The proofs are therefore as the proofs of Theorem 5.7 and Theorem 5.8 in Chapter 5 but with $S_{h,\Delta t}^{m-k}$ replaced by $S_h(t_m - t_k)$ and using the similar estimates as in the proofs of Theorem 6.1 and Theorem 6.3 for the SETD1 scheme.

6.4 Implementation & numerical results

As for deterministic exponential integrators, the key element in the stochastic exponential schemes is the computing of the matrix exponential functions, the so called φ_i-functions.

We use here the real fast Léja points and the Krylov subspace technique. Implementations of these two techniques are given in Chapter 3.

6.4.1 Numerical construction of noise

We relate the decay of the eigenvalues q_i of Q in (4.3) to the covariance function and discuss implementation. For concreteness we examine A on $[0, L_1]\times[0, L_2]$ with Neumann boundary conditions. For the process $O(t) \in L_2(\mathbb{D}, H^r(\Omega))$, $r = 1, 2$ we take the following values for $\{q_{i,j}\}_{i+j>0}$ in the representation (4.3)

$$q_{i,j} = \Gamma / \left(i + j\right)^{r/2}, \qquad r > 0. \tag{6.16}$$

We say that the noise is in H^r when the eigenvalues satisfy (6.16). Again we consider the covariance operator Q with the following covariance function (kernel) with strong exponential decay [75, 81, 92]

$$C_r((x_1, y_1); (x_2, y_2)) = \frac{\Gamma}{4b_1b_2} \exp\left(-\frac{\pi}{4}\left[\frac{\left(x_2 - x_1\right)^2}{b_1^2} + \frac{\left(y_2 - y_1\right)^2}{b_2^2}\right]\right)$$

where b_1, b_2 are spatial correlation lengths in the x and y directions respectively and $\Gamma > 0$. This covariance function is frequently used in geosciences to generate a random permeability (see [30] and Chapter 3). Recall that the corresponding values of $\{q_{i,j}\}_{i+j>0}$ in the representation (4.3) are given by

$$q_{i,j} = \Gamma \exp\left[-\frac{1}{2\pi}\left((\lambda_i^{(1)}b_1)^2 + (\lambda_j^{(2)}b_2)^2\right)\right],$$

During our simulation, the process

$$\widehat{O}_k = \int_{t_k}^{t_{k+1}} e^{(t_{k+1}-\tau)A_N} dW^N(\tau)$$

is generated in Fourier space as in [89] by applying the Ito isometry in each mode, which yields

$$(e_i, \widehat{O}_k)) = e^{-\lambda_i \Delta t}\left(\frac{q_i}{2\lambda_i}\left(1 - e^{-2\lambda_i\Delta t}\right)\right)^{1/2} R_{i,k}, \tag{6.17}$$

$i \in \mathcal{I}_N = \{1, 2, 3, ..., N\}^2$, $k = 0, 1, 2..., M - 1$ and $R_{i,k}$ are independent, standard normally distributed random variables with means 0 and variance 1. For efficient computations

we use the inverse fast Fourier transform or some variant : e.g. for Neumann boundary conditions we use the inverse discrete cosine transform.

The exponential functions in the schemes SETD0 and SETD1 are computed either using the real Léja points technique or the Krylov subspace technique. For noise with exponential correlations, $b_i > 0$, $i = 1, 2$ we have $\|(-A)^{r/2}Q^{1/2}\|_{HS} < \infty$, $r = 1, 2$. Furthermore Assumption 5.5 is obviously satisfied with $\mathcal{V} = H = L^2(\Omega)$ and $\theta = 1/2$. We therefore expect the higher temporal order, i.e. close to 1 with initial data $X_0 = 0$ when F is taken to be linear. We need to consider the projection P_h of the noise onto the computational grid. There are two cases. When the vertices of our finite element mesh matches the evaluation points of the noise term $O(t)$, the projection P_h is trivial. We also used the centered finite volume [35] discretization. Here P_h is trivial when the center of every control volume is an evaluation point of $O(t)$. Of course in general the evaluation points of the noise term $O(t)$ do not necessarily need to match the finite volume or finite element grids. In this case the noise needs to be regular for a good projection (see assumption 5.5).

In our simulations we examine both a finite element and a finite volume discretization in space and take as a domain $\Omega = [0, 1] \times [0, 1]$. For time discretizations we compare the schemes here with an semi-implicit Euler Maruyama method (denoted "Implicitfem") and the semi-implicit Euler Maruyama of Chapter 5 that uses linear functionals of the noise as in (6.17). We denote by "Implicitfem" the graph for the standard semi-implicit in time with finite element method for space discretization with exponential correlation function,"SETD1fem" and "SETD0fem" the graph for schemes SETD1 and SETD0 with finite element method for space discretization and exponential correlation function, "Implicitfvm$r$", $r = 1, 2$ the graph for standard implicit with finite volume method for space discretization with H^r noise, "SETD1fem$r$" and "SETD0fem$r$", $r = 1, 2$ the graph for the schemes SETD1 and SETD0 with finite element method for space discretization with H^r noise, "SETD1fvm$r$" and "SETD0fvm$r$", $r = 1, 2$ the graph for the schemes SETD1and SETD0 with finite volume method for space discretization with H^r noise, "ModifiedImplicitfvm$r$", $r = 1, 2$ graph for the modified implicit scheme constructed in Chapter 5 with finite volume method for space discretization with H^r noise.

6.4.2 A linear reaction–diffusion equation

As a simple example consider the reaction diffusion equation in the time interval $[0, T]$ with diffusion coefficient $D > 0$

$$dX = (D\Delta X - \lambda X)dt + dW \qquad X(0) = X_0, \tag{6.18}$$

with homogeneous Neumann boundary conditions in Ω. Here λ is a constant related to the reaction and in the notation of (5.1) $F(u) = -\lambda u$ and obviously satisfies condition (a) of Assumption (5.3). For this linear equation we can construct an exact solution up to any spectral projection error. We compute the exponential functions φ_i with the real fast Léja points technique. The absolute tolerance used is 10^{-6}. As in Chapter 5, notice that $A = D\Delta$ does not satisfy Assumption 5.2 as 0 is an eigenvalue. During the simulations we need to manage the singularity of (6.17) at $\lambda_0 = 0$ or use the perturbed operator $A = D\Delta + \epsilon\mathbf{I}$, $\epsilon > 0$.

We start by examining in Figure 6.1 convergence with H^r noise, $r = 1, 2$. The figure compares the finite element discretization for schemes SETD0, SETD1, the standard implicit Euler–Maruyama scheme and the modified implicit scheme introduced in [18] which also uses a linear functional of the noise. We observe that schemes with finite element and finite volume space discretization have the same order of accuracy. In Figure 6.1 (a) the noise is in H^1 and the diffusion coefficient is $D = 1$. We clearly see improved accuracy of the schemes that use the linear functions of the noise : namely SETD0, SETD1 and modified implicit over the standard semi-implicit method. Not only is there an improved constant but the temporal order is higher. Numerically we find from Figure 6.1 an order of 0.97 for SETD0, SETD1 and for the modified semi-implicit Euler-Maruyama scheme, which are in excellent agreement with the theoretical value of 1 from the theory, the order of convergence of the standard implicit scheme is 0.30. We also see that the scheme SETD0 and the modified implicit scheme have approximately the same order of accuracy and that SETD1 is slightly more accurate compared to the schemes SETD0 and the modified semi-implicit Euler-Maruyama. In Figure 6.1 (b) the noise is H^2 and diffusion coefficient $D = 1/100$. The error here is dominated by space discretization error, as a consequence to see the convergence we need small Δx and Δy. We observe again that the schemes using the linear functionals are more accurate. We also see from both Figure 6.1 (a) and (b) that SETD1 is

slightly more accurate than SETD0 by some constant. The temporal order of convergence for schemes using linear functional of the noise is 0.97 and 0.5 for the standard semi-implicit scheme. From Figure 6.1 (a) to Figure 6.1 (b) we observe that as the noise is regular the gap between errors in different schemes is small.

In Figure 6.2 we show results with the exponential covariance function for the noise, as the noise is certainly in $H^r,\ r = 1$ or 2 we expect a rate of convergence close to one. The figure compares the finite element discretization for schemes SETD0 and SETD1 against the standard implicit scheme. The temporal order of convergence of the schemes SETD0 is 0.80 and SETD1 is 1.05 and 0.80 for the standard implicit scheme. We see the improved accuracy in the schemes SETD0 and SETD1 compared to the standard implicit scheme. We also see the better accuracy of the scheme SETD1 compared to SETD0.

6.4.3 Stochastic advection diffusion reaction

As a more challenging example we consider the stochastic advection diffusion reaction SPDE

$$dX = \left(D\Delta X - \nabla \cdot (\mathbf{q}X) - \frac{X}{X+1} \right) dt + dW, \tag{6.19}$$

with mixed Neumann-Dirichlet boundary conditions. and constant velocity $\mathbf{q} = (1, 0)$ for homogeneous medium. The Dirichlet boundary condition is $X = 1$ at $x = 0$ and we use the homogeneous Neumann boundary conditions elsewhere. In terms of equation (5.1) the nonlinear term F is given by

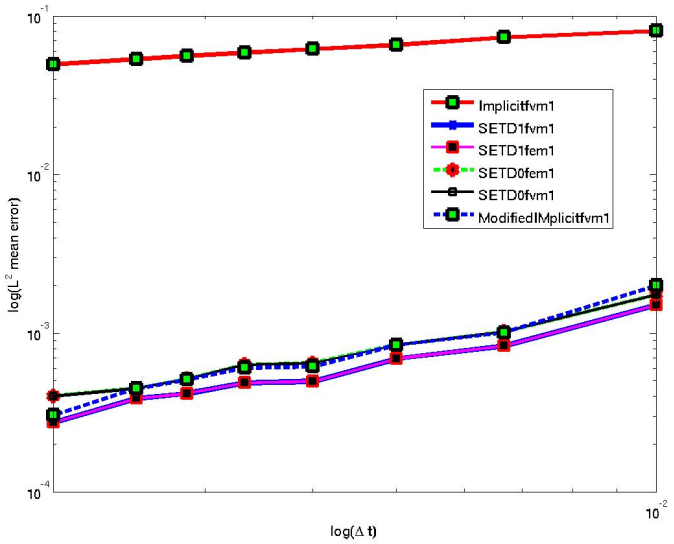
$$F(u) = -\nabla \cdot (\mathbf{q}u) - \frac{u}{(u+1)}, \quad u \in \mathbb{R}^+ \tag{6.20}$$

and clearly satisfies Assumption 5.3 (b). In our simulation the intensity of the noise is such that we cannot have $X = -1$. We can also bypass the singularity issue by taking

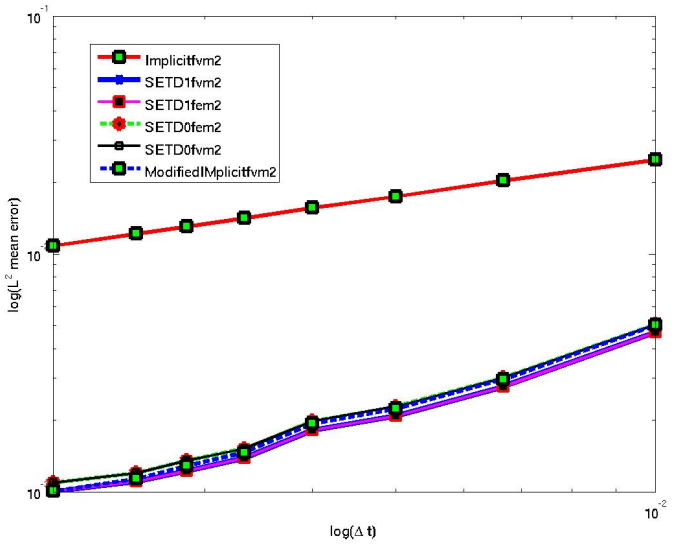
$$F(u) = -\nabla \cdot (\mathbf{q}u) - \frac{u}{(|u|+1)}, \quad u \in \mathbb{R} \tag{6.21}$$

as in the Chapter 5. For heterogeneous a medium we used three parallel high permeability streaks. This could represent for example a highly idealized fracture pattern. We obtain the Darcy velocity field $\mathbf{q}$ by solving the system

$$\begin{cases} \nabla \cdot \mathbf{q} = 0 \\ \mathbf{q} = -\dfrac{k(\mathbf{x})}{\mu} \nabla p, \end{cases} \tag{6.22}$$



(a)



(b)

Figure 6.1: Convergence in the root mean square L^2 norm at $T = 1$ as a function of Δt with H^r, $r = 1, 2$. (a) Shows convergence for finite element and finite volume discretizations with $r = 1$, $D = 1$, $\lambda = 1$, $\Gamma = 1$ and $\Delta x = \Delta y = 1/100$. In (b) we show convergence for finite element and finite volume discretizations with $r = 2$, $D = 1/100$, $\lambda = 1$, $\Gamma = 1$, $\Delta x = \Delta y = 1/400$ (small to have a good look of convergence). The initial data is $X_0 = 0$ and the simulation is for (6.18) with 20 realizations.

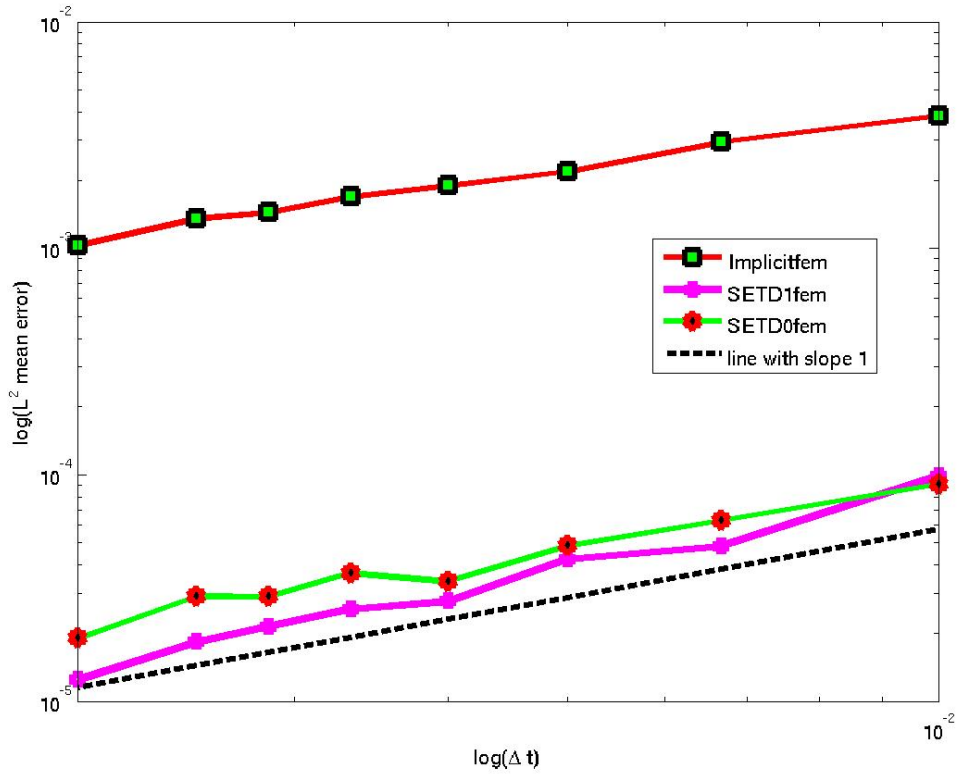


Figure 6.2: Convergence in the root mean square L^2 norm at $T = 1$ as a function of Δt with exponential covariance function with $D = 1$, $\lambda = 0.5$, $\Gamma = 1$ and regular mesh coming from rectangular grid with size $\Delta x = \Delta y = 1/100$. The simulation is for (6.18) with correlation lengths $b_1 = b_2 = 0.2$ and 10 realizations. Initial data is given by $X_0 = 0$.

with Dirichlet boundary conditions $\Gamma_D^1 = \{0,1\} \times [0,1]$ and Neumann boundary $\Gamma_N^1 = (0,1) \times \{0,1\}$ such that

$$p = \begin{cases} 1 & \text{in} \quad \{0\} \times [0,1] \\ 0 & \text{in} \quad \{L_1\} \times [0,1] \end{cases}$$

and

$$-k\,\nabla p(\mathbf{x},t) \cdot \mathbf{n} = 0 \quad \text{in} \quad \Gamma_N^1$$

where p is the pressure, μ is dynamical viscosity and k the permeability of the porous medium. We have assumed that rock and fluids are incompressible and sources or sinks are absent, thus the equation

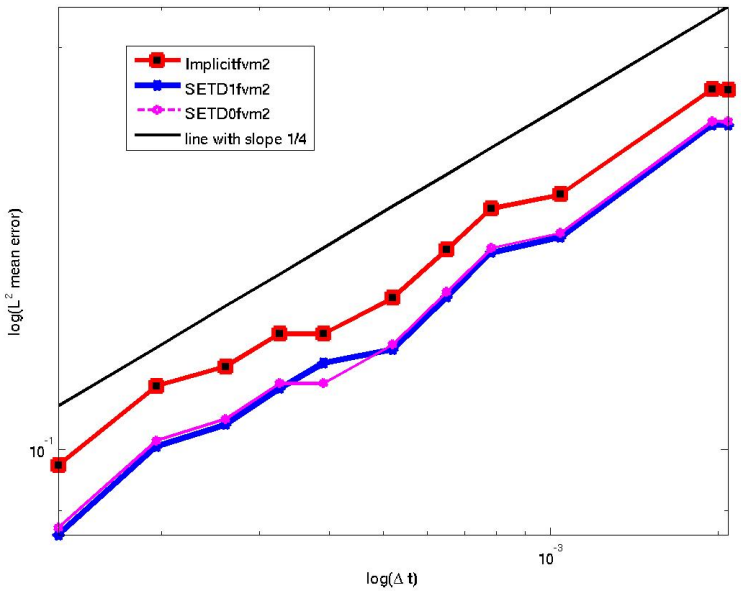
$$\nabla \cdot \mathbf{q} = \nabla \cdot \left[\frac{k(\mathbf{x})}{\mu} \nabla p \right] = 0 \tag{6.23}$$

comes from mass conservation. To deal with high Péclet flows we discretize in space using finite volumes. Simulations are in $L^2(\Omega)$ since the discrete $L^2(\Omega)$ norm is easy to implement for all types of boundary conditions. We can write the semi-discrete finite volume method as

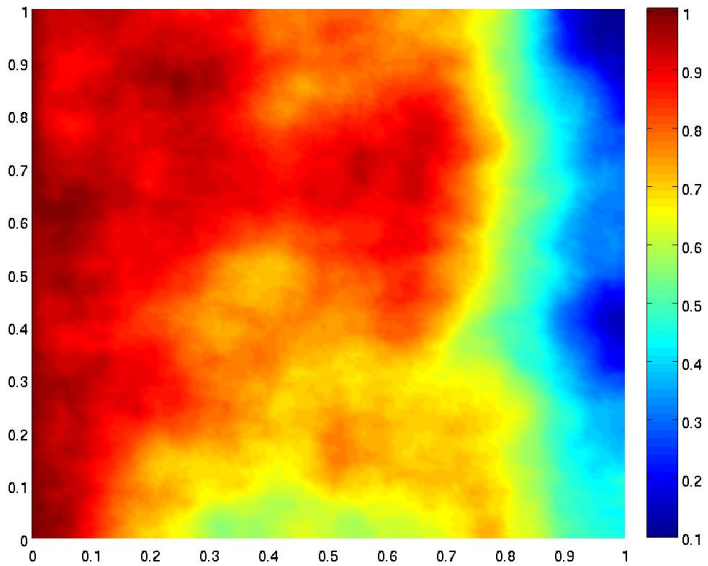
$$dX^h = (A_h X^h + P_h F(X^h) + b(X^h)) + P_h P_N dW, \tag{6.24}$$

where here A_h is the space discretization of $D\Delta$ using only homogeneous Neumann boundary conditions and $b(X^h)$ comes from the approximation of diffusion flux at the Dirichlet boundary condition size. We compute the exponential functions φ_i with Krylov subspace technique with dimension $m = 6$ and the absolute tolerance 10^{-6} and the real fast Léja points technique for φ_0. In Figure 6.3(a) we show the convergence of schemes SETD0, SETD1 and standard implicit scheme with H^2 noise for homogeneous medium, the "true solution" is the numerical scheme with smaller time step $\Delta t = 1/15360$. All the schemes have temporal order of convergence 1/4. We can also observe the accuracy of the scheme SETD1 and SETD0 comparing to the standard implicit scheme in Figure 6.3(a). In Figure 6.4(a) we show the convergence of schemes SETD0, SETD1 with H^2 noise for a heterogeneous medium. The two schemes have the same error. The corresponding mean of CPUtime for the scheme SETD0 is given in Figure 6.5(b). We observe a slight efficiency gain using the Léja points technique compared to the Krylov subspace technique during the evaluation of the action of φ_0.

In conclusion we obtained superior convergence for the stochastic exponential integrators using linear functionals of the noise with a finite element discretization. Furthermore we have shown that these schemes that require the exponential of a non-diagonal matrix can be efficiently implemented for finite element and finite volume discretizations of realistic porous media flow with stochastic forcing.

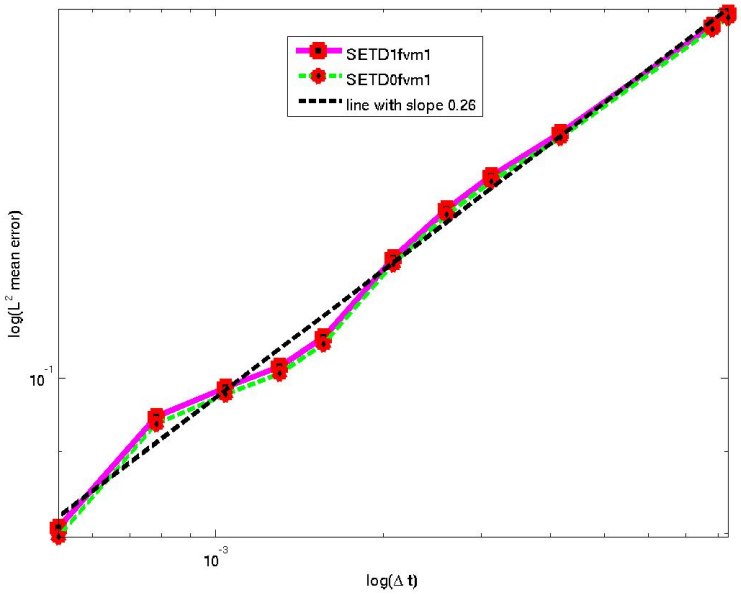


(a)

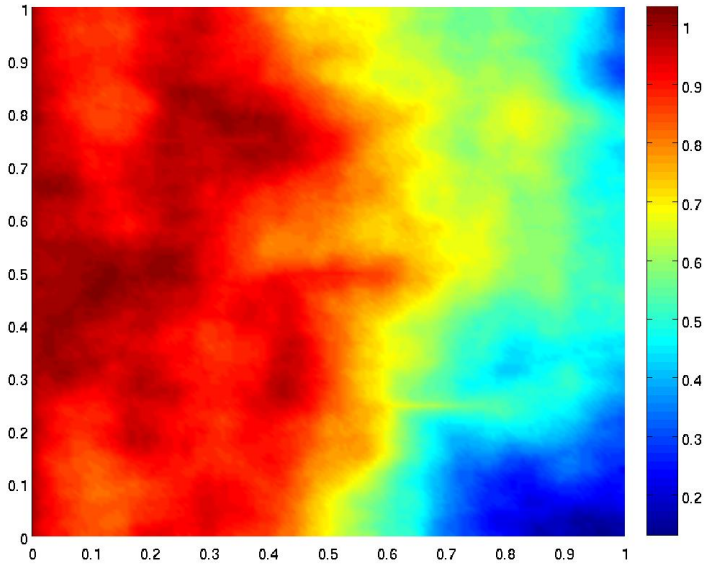


(b)

Figure 6.3: (a) Convergence of the root mean square L^2 norm at $T = 1$ as a function of Δt with 30 realizations with $\Delta x = \Delta y = 1/160$, $X_0 = 0$, $\Gamma = 0.01$ for homogeneous medium. The noise is white in time and in H^2 in space. The temporal order of convergence in time is 1/4 for all schemes. In (b) we plot a sample of a "true solution" for $r = 2$ with $\Delta t = 1/15360$.

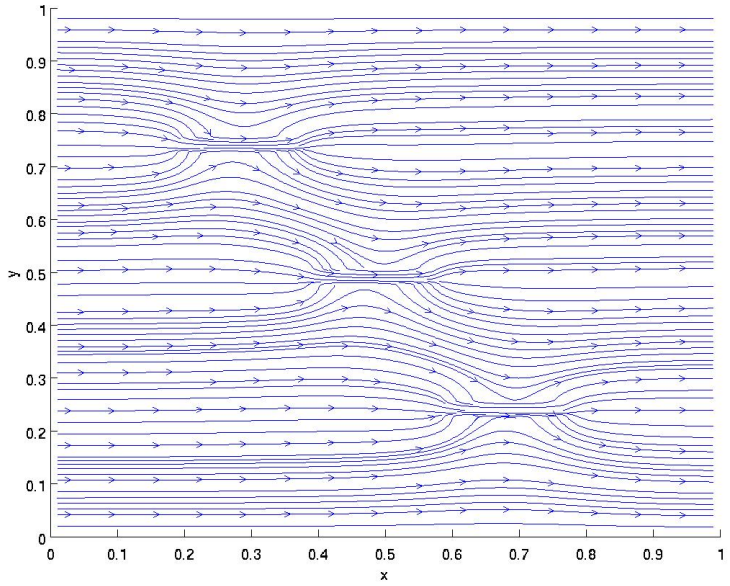


(a)

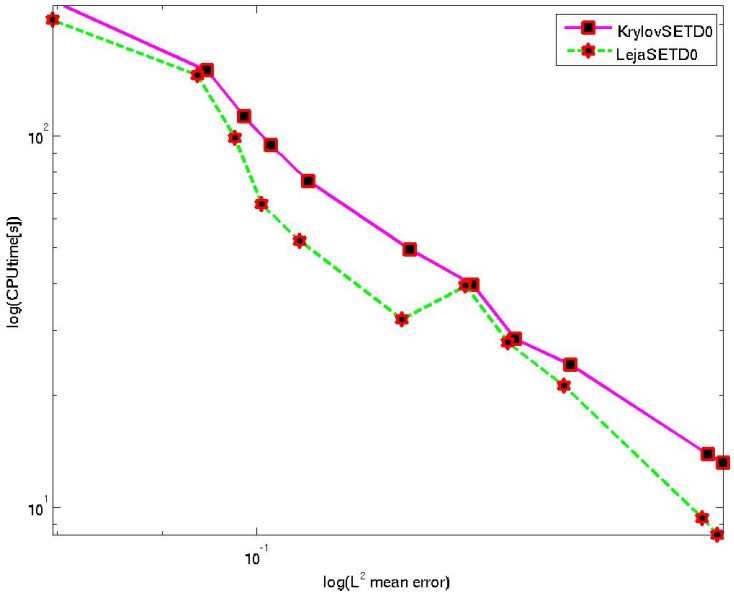


(b)

Figure 6.4: (a) Convergence of the root mean square L^2 norm at $T = 1$ as a function of Δt with 30 realizations and $\Delta x = \Delta y = 1/160$, $X_0 = 0$, $\Gamma = 0.01$ for heterogeneous medium. The noise is white in time and in H^2 in space. The temporal order of convergence in time is 0.26 (close to 1/4) for the two methods. In (b) we plot a sample of a "true solution" for $r = 1$ with $\Delta t = 1/15360$.



(a)



(b)

Figure 6.5: Streamline of the velocity in heterogeneous medium with comparison Krylov subspace and real and Léja points techniques corresponding to Figure 6.4. In (a) we plot the streamline of the velocity field and in (b) the mean CPUtime for SETD0 using the Krylov and Leja points techniques. The local Péclet number for the flow is 16.58.

Chapter 7

Stochastic Exponential Integrators for Finite Element Discretization of SPDEs for Multiplicative & Additive Noise

In this chapter, we consider the numerical approximation of a general second order semi–linear parabolic stochastic partial differential equation (SPDEs) driven by space-time noise, for additive and multiplicative noise. In contrast to the standard time stepping methods which approximate the exponential operator by the rational fraction operators in the mild form, we extend two deterministic exponential integrators to stochastic exponential integrators schemes.

We consider noise that is in the trace class and give a convergence proof in the mean square L^2 norm. We discretize in space with the finite element method and in our implementation we examine both the finite element and the finite volume methods. In our numerical results we use two different efficient techniques to compute the exponential matrix functions: the real fast Léja points and the Krylov subspace techniques as in Chapter 6. We present results for a linear reaction diffusion equation in two dimensions with additive noise as well as a nonlinear example of two-dimensional stochastic advection diffusion reaction equation motivated from realistic porous media flow. Results from this chapter are presented in our paper [94].

7.1 Introduction

We analyse the strong numerical approximation of the Ito stochastic partial differential equation defined in $\Omega \subset \mathbb{R}^d$. Boundary conditions on the domain Ω are typically Neumann, Dirichlet or some mixed conditions. We consider the general SPDEs given in (4.11) with $H = L^2(\Omega)$. Typical examples are stochastic (advection) diffusion reaction equations where $A = \nabla \cdot (\mathbf{D}\nabla(.))$ arising from example in pattern formation in physics and mathematical biology. We illustrate our work with both a simple reaction diffusion equation where we can construct an exact solution

$$dX = (\nabla \cdot (\mathbf{D}\nabla X) - \lambda X)\, dt + dW \tag{7.1}$$

as well as the stochastic advection reaction diffusion equation

$$dX = \left(\nabla \cdot (\mathbf{D}\nabla X) - \nabla \cdot (\mathbf{q}X) - \frac{X}{|X| + 1}\right) dt + X dW \tag{7.2}$$

where $\mathbf{D}$ is the diffusion matrix, $\mathbf{q}$ is the Darcy velocity field [20] and λ is a constant depending of the reaction function.

Our schemes here are based on using the finite element method (or finite volume method) for space discretization so that we gain the flexibility of these methods to deal with complex boundary conditions and we can apply well developed techniques such as upwinding to deal with advection. We improve on the schemes presented in Chapter 5 and Chapter 6 as we do not require the linear operator A to be self adjoint and do not need information on the eigenvalues and eigenfunctions of the operator A.

As in Chapter 6, schemes presented here are based on exponential matrix computation where Léja points and Krylov subspace techniques are efficient tools. The convergence proof given below is similar to one in [87]. The chapter is organised as follows. In Section 7.2 we present the two numerical schemes based on the exponential integrators and our assumptions on (4.11). We also present and comment on our convergence results. Section 7.3 contains the proofs of our convergence theorems. We conclude in Section 7.4 by presenting some simulations and discuss implementation of these methods.

7.2 Numerical schemes and main results

7.2.1 The abstract setting

We assume that Ω has a smooth boundary or is a convex polygon of $\mathbb{R}^d,\ d = 1, 2, 3$. For convenience of presentation we take A to be a second order operator as this simplifies the convergence proof. More precisely we consider the general second order semi–linear parabolic stochastic partial differential equation given by

$$\begin{aligned} dX(t,\mathbf{x}) &= (\nabla \cdot \mathbf{D}\nabla X(t,\mathbf{x}) - \mathbf{q} \cdot \nabla X(t,\mathbf{x}) + f(\mathbf{x}, X(t,\mathbf{x}))) \, dt \\ &\quad + b(\mathbf{x}, X(t,\mathbf{x}))dW(t,\mathbf{x}), \qquad \mathbf{x} \in \Omega, \ \ t \in [0,T], \end{aligned} \tag{7.3}$$

where $f, b : \Omega \times \mathbb{R} \to \mathbb{R}$ are two continuously differentiable functions with globally bounded derivatives.

In the abstract form given in (4.11), the linear operator is defined by

$$\begin{aligned} A &= \nabla \cdot \mathbf{D}\nabla(.) - \mathbf{q} \cdot \nabla(.) \\ &= \sum_{i,j=1}^{d} \frac{\partial}{\partial x_i} \left(D_{i,j} \frac{\partial}{\partial x_j} \right) - \sum_{i=1}^{d} q_i \frac{\partial}{\partial x_i}, \end{aligned} \tag{7.4}$$

where we assume as in Chapter 2 that $D_{i,j} \in L^\infty(\Omega)$, $q_i \in L^\infty(\Omega)$ and that there exists a positive constant $c_1 > 0$ such that (2.14) holds. The nonlinear operators $F : H \to H$ and $B : H \to HS(Q^{1/2}(H), H) = L_2^0$ are defined by

$$(F(v))(\mathbf{x}) = f(\mathbf{x}, v(\mathbf{x})), \qquad (B(v)u)(\mathbf{x}) = b(\mathbf{x}, v(\mathbf{x})) \cdot u(\mathbf{x}), \tag{7.5}$$

for all $\mathbf{x} \in \Omega,\ v \in H,\ u \in Q^{1/2}(H)$, with $H = L^2(\Omega)$.

According to Chapter 2, the linear operator A generate an analytic semigroup $S(t) = e^{tA}$. The nonlinear operators $F,\ B$ and the noise need to satisfy Assumption 4.12, Assumption 4.13 (or Assumption 4.14) for existence and uniqueness of the mild solution of equation (4.11). Notice that by the definitions of the operator B and $\|.\|_{L_2^0}$, for $Y \in H = L^2(\Omega)$ we have

$$\|B(Y)\|_{L_2^0}^2 = \sum_{i \in \mathbb{N}^d} \|b(Y)Q^{1/2}e_i\|^2, \tag{7.6}$$

where $b(Y)$ is the Nemytskii operator defined by

$$b(Y)(\mathbf{x}) = b(\mathbf{x}, Y(\mathbf{x})) \qquad \mathbf{x} \in \Omega. \tag{7.7}$$

The following theorem gives the regularity result of the mild solution X of (4.11).

Theorem 7.1 *Assume that Assumption 4.12 and Assumption 4.13 hold. Let X be the mild solution given in (4.13). If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\beta/2}))$, $\beta \in [0,1)$ then for all $t \in [0,T]$, $X(t) \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\beta/2}))$ with*

$$\left(\mathbf{E}\|X(t)\|_\beta^2\right)^{1/2} \leq C\left(1 + \left(\mathbf{E}\|X_0\|_\beta^2\right)^{1/2} + \sup_{0\leq s\leq t}\left(\mathbf{E}\|X(s)\|^2\right)^{1/2}\right).$$

Proof. Recall that if X is the solution of (4.11), it therefore satisfies the boundary condition and we only need to check that

$$\left(\mathbf{E}\|X(t)\|_\beta^2\right)^{1/2} < \infty$$

to conclude the proof. Recall that the mild solution is given by

$$X(t) = S(t)X_0 + \int_0^t S(t-s)F(X(s))ds + \int_0^t S(t-s)B(X(s))dW(s)$$

then

$$\begin{aligned}
\left(\mathbf{E}\|X(t)\|_\beta^2\right)^{1/2} &\leq \left(\mathbf{E}\|S(t)X_0\|_\beta^2\right)^{1/2} + \left(\mathbf{E}\|\int_0^t S(t-s)F(X(s))ds\|_\beta^2\right)^{1/2} \\
&\quad + \left(\mathbf{E}\|\int_0^t S(t-s)B(X(s))dW(s)\|_\beta^2\right)^{1/2} \\
&= I + II + III.
\end{aligned}$$

We obviouly have

$$I = \left(\mathbf{E}\|S(t)X_0\|_\beta^2\right)^{1/2} \leq C\left(\mathbf{E}\|X_0\|_\beta^2\right)^{1/2}.$$

Let us estimate II

$$\begin{aligned}
II &= \left(\mathbf{E}\|\int_0^t S(t-s)F(X(s))ds\|_\beta^2\right)^{1/2} \\
&\leq \int_0^t \left(\mathbf{E}\|S(t-s)F(X(s))\|_\beta^2\right)^{1/2} ds \\
&\leq \int_0^t \left(\mathbf{E}\|(-A)^{\beta/2}S(t-s)F(X(s))\|^2\right)^{1/2} ds.
\end{aligned}$$

Using the consequence of Assumption 4.12 and the semigroup properties in Proposition 2.6 yields

$$\begin{aligned} II &\leq C\left(\int_0^t \|(-A)^{\beta/2}S(t-s)\|_{L(L^2(\Omega))}ds\right)\left(\sup_{0\leq s\leq t}\left(\mathbf{E}\left(1+\|X(s)\|\right)^2\right)^{1/2}\right) \\ &\leq C\left(\int_0^t (t-s)^{-\frac{\beta}{2}}ds\right)\left(1+\sup_{0\leq s\leq t}\left(\mathbf{E}\|X(s)\|^2\right)^{1/2}\right) \\ &\leq C\left(1+\sup_{0\leq s\leq t}\left(\mathbf{E}\|X(s)\|^2\right)^{1/2}\right). \end{aligned}$$

Finally, Ito's isometry and the consequence of Assumption 4.13 yields

$$\begin{aligned} III^2 &= \mathbf{E}\|\int_0^t S(t-s)B(X(s))dW(s)\|_\beta^2 \\ &= \int_0^t \mathbf{E}\|(-A)^{\beta/2}S(t-s)B(X(s))\|_{L_2^0}^2 ds \\ &\leq C\left(\int_0^t \|(-A)^{\beta/2}S(t-s)\|_{L(L^2(\Omega))}^2 ds\right)\left(1+\sup_{0\leq s\leq t}\mathbf{E}\|X(s)\|^2\right) \\ &\leq C\left(1+\sup_{0\leq s\leq t}\mathbf{E}\|X(s)\|^2\right), \end{aligned}$$

thus

$$III \leq C\left(1+\sup_{0\leq s\leq t}\left(\mathbf{E}\|X(s)\|^2\right)^{1/2}\right).$$

Then

$$\left(\mathbf{E}\|X(t)\|_\beta^2\right)^{1/2} \leq C\left(1+\left(\mathbf{E}\|X_0\|_\beta^2\right)^{1/2}+\sup_{0\leq s\leq t}\left(\mathbf{E}\|X(s)\|^2\right)^{1/2}\right) < \infty.$$

■

More results about the regularity of X can be found in [95, 96]. For the same initial data, [95] extends the regularity of the solution X.

7.2.2 Numerical schemes

As in the previous chapter, we consider the discretization of the spatial domain by a finite element triangulation $\mathcal{T}_h$.

The semi–discrete in space version of the problem (4.11) is to find the process $X^h(t) = X^h(.,t) \in V_h$ such that for $t \in [0,T]$,

$$dX^h = (A_h X^h + P_h F(X^h))dt + P_h B(X^h)dW, \qquad X^h(0) = P_h X_0. \tag{7.8}$$

The mild solution of (7.8) at time $t_m = m\Delta t,\ \ \Delta t > 0\ \ $ is given by

$$\begin{aligned} X^h(t_m) &= S_h(t_m)P_h X_0 + \int_0^{t_m} S_h(t_m - s)P_h F(X^h(s))ds \\ &\quad + \int_0^{t_m} S_h(t_m - s)P_h B(X^h)dW(s). \end{aligned} \tag{7.9}$$

Then, given the mild solution at the time t_m, we can construct the corresponding solution at t_{m+1} as

$$\begin{aligned} X^h(t_{m+1}) &= S_h(\Delta t)X^h(t_m) + \int_0^{\Delta t} S_h(\Delta t - s)P_h F(X^h(s + t_m))ds \\ &\quad + \int_{t_m}^{t_{m+1}} S_h(t_{m+1} - s)P_h B(X^h)dW(s). \end{aligned}$$

To build the first numerical scheme, we use the following approximations

$$\begin{aligned} S_h(\Delta t - s)F(X^h(t_m + s)) &\approx S_h(\Delta t)F(X^h(t_m) \qquad s \in [0,\ \Delta t], \\ S_h(t_{m+1} - s)P_h B(X^h) &\approx S_h(\Delta t)P_h B(X^h(t_m)) \quad s \in [t_m, t_{m+1}]. \end{aligned}$$

We can define our approximation Y_m^h of $X(m\Delta t)$ by

$$\begin{aligned} Y_{m+1}^h &= e^{\Delta t A_h}\left(Y_m^h + P_h F(Y_m^h) + P_h B(Y_m^h)\left(W_{m+1} - W_m\right)\right) \\ &= \varphi_0(\Delta t A_h)\left(Y_m^h + P_h F(Y_m^h) + P_h B(Y_m^h)\Delta W_m\right), \end{aligned} \tag{7.10}$$

where

$$\begin{aligned} \varphi_0(\Delta t A_h) &= e^{\Delta t A_h} \\ \Delta W_m &= W_{m+1} - W_m = \sqrt{\Delta t}\sum_{i \in \mathbb{N}^d} \sqrt{q_i} R_{i,m} e_i, \end{aligned}$$

with $R_{i,m}$ are independent, standard normally distributed random variables with means 0 and variance 1. We call the scheme defined by (7.10) SETDM0. In order to build the

second numerical scheme, we use the following approximations

$$F(X^h(t_m+s)) \approx F(X^h(t_m)) \qquad s\in[0,\ \Delta t],$$

$$S_h(t_{m+1}-s)P_hB(X^h) \approx S_h(\Delta t)P_hB(X^h(t_m)) \quad s\in[t_m,t_{m+1}].$$

We can define our approximation X_m^h of $X(m\Delta t)$ by

$$X_{m+1}^h = e^{\Delta t A_h}X_m^h + A_h^{-1}\left(e^{\Delta t A_h}-I\right)P_hF(X_m^h) + e^{\Delta t A_h}P_hB(X_m^h)\left(W_{m+1}-W_m\right). \quad (7.11)$$

For efficiency we rewrite the scheme (7.11) as

$$X_{m+1}^h = X_m^h + \Delta t\varphi_1(\Delta t A_h)\left(A_h\left(X_m^h + P_hB(X_m^h)\Delta W_m\right) + P_hF(X_m^h)\right),$$

where

$$\varphi_1(\Delta t A_h) = (\Delta t\, A_h)^{-1}\left(e^{\Delta t A_h}-I\right) = \frac{1}{\Delta t}\int_0^{\Delta t} e^{(\Delta t-s)A_h}ds.$$

We will call this second scheme SETDM1. This scheme is also used in [97] with the Fourier method to solve fourth order stochastic problems.

7.2.3 Main result

Throughout the chapter we take $t_m = m\Delta t \in (0,T]$, where $T = M\Delta t$ for $m, M \in \mathbb{N}$. We take C to be a constant that may depend on T and other parameters but not on Δt or h.

Our main result is a strong convergence result in L^2 for schemes SETDM1 and SETDM0.

Theorem 7.2 *Suppose Assumption 4.12, Assumption 4.13 (or Assumption 4.14) are satisfied. Let $X(t_m)$ be the mild solution of equation (4.11) represented by (4.13) and suppose that $b(X(t)) \in L_2(\mathbb{D}, \mathcal{D}((-A)^\alpha))$, $\forall t\in[0,T]$, $\alpha\in(0,1/2)$ small enough. Let ζ_m^h be the numerical approximation through scheme (7.11) or (7.10) ($\zeta_m^h = X_m^h$ for scheme SETDM1 and $\zeta_m^h = Y_m^h$ for scheme SETDM0) and $0<\gamma<1$. The following estimates hold: If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $0<\gamma\le 1/2$ then*

$$\left(\boldsymbol{E}\|X(t_m)-\zeta_m^h\|^2\right)^{1/2} \le C\left(t_m^{(-1+2\gamma)/2}h + \Delta t^{\gamma/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $1/2\le\gamma<1$ then

$$\left(\boldsymbol{E}\|X(t_m)-\zeta_m^h\|^2\right)^{1/2} \le C\left(h + \Delta t^{\gamma/2}\right).$$

Suppose that for $t \in [0,T]$, $F(X(t)) \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\beta})), \forall t \in [0,T]$, $\beta \in (0,1/2)$ *small enough: If* $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\gamma}))$, *for* $0 < \gamma < 1$ *then*

$$\left(\boldsymbol{E}\|X(t_m) - \zeta_m^h\|^2\right)^{1/2} \leq C\left(t_m^{(-1+\gamma)}h^2 + \Delta t^{\gamma/2}\right).$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ *and* $b(X(t)) \in L_2(\mathbb{D}, \mathcal{D}(-A))$, $\forall t \in [0,T]$ *then*

$$\left(\boldsymbol{E}\|X(t_m) - \zeta_m^h\|^2\right)^{1/2} \leq C\left(h^2 + \Delta t^{(1/2-\epsilon)}\right).$$

$\epsilon \in (0,1/2)$, *small enough.*

Similar results hold, however, for higher order operators.

Computationally, the noise given by (4.3) is truncated to N terms. Therefore the corresponding approximated solutions become $X_m^{h,N}$ for SETDM1 and $Y_m^{h,N}$ for scheme SETDM0. For noise where the eigenvalues of the covariance operator has strong exponential decay, $X_m^{h,N}$ and $Y_m^{h,N}$ are close to X_m^h and Y_m^h respectively. In the case of additive noise, it has been proved in [78] that with the truncation to N terms of the noise (4.3) the corresponding discrete mild solution $X^{h,N}$ in (7.9) has the same order of accuracy with respect to h as X^h.

Note that weak assumption on the noise improve the accuracy in Theorem 7.2 for additive noise (see [91] and simulation in Section 7.4.1).

7.3 Proofs of main results

7.3.1 Preparatory result

Our preliminary lemma concerns the mild solution of SPDE (4.11).

Lemma 7.3 *Let* X *be the mild solution given in (4.13). Suppose that Assumption 4.12 holds on* F *and Assumption 4.13 holds for* B. *Let* $0 \leq \gamma < 1$, $t_1, t_2 \in [0,T]$ *be so that* $t_1 < t_2$. *If* $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\gamma}))$ *then we have the following estimate,*

$$\begin{aligned}
&\boldsymbol{E}\|X(t_2) - X(t_1)\|^2 \\
&\leq C(t_2 - t_1)^{\gamma}\left(\boldsymbol{E}\|X_0\|_{\gamma}^2 + \boldsymbol{E}\left(\sup_{0\leq s\leq T}(1 + \|X(s)\|)\right)^2 + \sup_{0\leq s\leq t_1}(1 + \boldsymbol{E}\|X(s)\|)^2\right).
\end{aligned}$$

Proof. Consider the difference

$$\begin{aligned} & X(t_2) - X(t_1) \\ = \; & \left(S(t_2) - S(t_1)\right) X_0 + \left(\int_0^{t_2} S(t_2 - s) F(X(s)) ds - \int_0^{t_1} S(t_1 - s) F(X(s)) ds \right) \\ & + \left(\int_0^{t_2} S(t_2 - s) B(X) dW(s) - \int_0^{t_1} S(t_1 - s) B(X) dW(s) \right) \\ = \; & I + II + III \end{aligned}$$

so that $\mathbf{E}\|X(t_2) - X(t_1)\|^2 \leq 3(\mathbf{E}\|I\|^2 + \mathbf{E}\|II\|^2 + \mathbf{E}\|III\|^2)$. We estimate each of the terms I, II and III. Estimation of the terms I and II are similar to ones in Chapter 5, Lemma 5.10 with additive noise. Using Proposition 2.6 as in Chapter 5 yields

$$\|I\| \;=\; \|S(t_1)(-A)^{-\gamma/2}(\mathrm{I} - S(t_2 - t_1))(-A)^{\gamma/2} X_0\| \;\leq\; C(t_2 - t_1)^{\gamma/2} \|X_0\|_\gamma$$

and

$$\mathbf{E}\|II\|^2 \leq C(t_2 - t_1)^{2\gamma} \mathbf{E} \left(\sup_{0 \leq s \leq T} (1 + \|X(s)\|) \right)^2.$$

For term III, we have

$$\begin{aligned} III \;&=\; \int_0^{t_1} \left(S(t_2 - s) - S(t_1 - s)\right) B(X) dW(s) + \int_{t_1}^{t_2} S(t_2 - s) B(X) dW(s) \\ &=\; III_1 + III_2. \end{aligned}$$

Using the Ito isometry property

$$\begin{aligned} \mathbf{E}\|III_1\|^2 \;&=\; \mathbf{E}\| \int_0^{t_1} \left(S(t_2 - s) - S(t_1 - s)\right) B(X) dW(s)\|^2 \\ &=\; \int_0^{t_1} \mathbf{E}\| \left(S(t_2 - s) - S(t_1 - s)\right) B(X)\|_{L_2^0}^2 ds. \end{aligned}$$

For $0 \leq \gamma < 1$, using Assumption 4.13 and Proposition 2.6 yields

$$
\begin{aligned}
&\mathbf{E}\|III_1\|^2 \\
&\leq C\left(\int_0^{t_1} \|(S(t_2-s)-S(t_1-s))\|^2_{L(L^2(\Omega))}ds\right)\left(\sup_{0\leq s\leq t_1}\mathbf{E}\left(1+\|X(s)\|\right)^2\right) \\
&= C\left(\int_0^{t_1} \|S(t_1-s)(-A)^{\gamma/2}(-A)^{-\gamma/2}(\mathbf{I}-S(t_2-t_1))\|^2_{L(L^2(\Omega))}ds\right)\left(\sup_{0\leq s\leq t_1}\mathbf{E}\left(1+\|X(s)\|\right)^2\right) \\
&= C\left(\int_0^{t_1} \|(-A)^{\gamma/2}S(t_1-s)(-A)^{-\gamma/2}(\mathbf{I}-S(t_2-t_1))\|^2_{L(L^2(\Omega))}ds\right)\left(\sup_{0\leq s\leq t_1}\mathbf{E}\left(1+\|X(s)\|\right)^2\right) \\
&\leq C(t_2-t_1)^{\gamma}\left(\int_0^{t_1}(t_1-s)^{-\gamma}ds\right)\left(\sup_{0\leq s\leq t_1}\mathbf{E}\left(1+\|X(s)\|\right)^2\right) \\
&\leq C(t_2-t_1)^{\gamma}\left(\sup_{0\leq s\leq t_1}\mathbf{E}\left(1+\|X(s)\|\right)^2\right).
\end{aligned}
$$

Let us estimate $\mathbf{E}\|III_2\|$. The Ito isometry, boundedness of S and Assumption 4.13 yield

$$
\begin{aligned}
\mathbf{E}\|III_2\|^2 &= \mathbf{E}\|\int_{t_1}^{t_2} S(t_2-s))B(X)dW(s)\|^2 \\
&= \int_{t_1}^{t_2}\mathbf{E}\|S(t_2-s))B(X(s))\|^2_{L_2^0}ds \\
&= \left(\int_{t_1}^{t_2}\|S(t_2-s)\|^2_{L(L^2(\Omega))}\,ds\right)\left(\sup_{0\leq s\leq t_1}\mathbf{E}\left(1+\|X(s)\|\right)^2\right) \\
&\leq C(t_2-t_1)\left(\sup_{0\leq s\leq t_1}\mathbf{E}\left(1+\|X(s)\|\right)^2\right).
\end{aligned}
$$

Hence

$$\mathbf{E}\|III\|^2 \leq 2(\mathbf{E}\|III_1\|^2+\mathbf{E}\|III_2\|^2) \leq C(t_2-t_1)^{\gamma}.$$

Combining our estimates of $\mathbf{E}\|I\|^2, \mathbf{E}\|II\|^2$ and $\mathbf{E}\|III\|^2$ ends the proof. ■

7.3.2 Proof of Theorem 7.2 for the scheme SETDM1

Proof. Let us rewrite the mild solution X as

$$
\begin{aligned}
X(t_m) &= S(t_m)X_0+\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S(t_m-s)F(X(s))ds+\int_0^{t_m} S(t_m-s)B(X(t_m))dW(s) \\
&= \overline{X}(t_m)+O(t_m),
\end{aligned}
$$

with

$$\begin{aligned}\overline{X}(t_m) &= S(t_m)X_0 + \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S(t_m - s)F(X(s))ds \\ O(t_m) &= \int_0^{t_m} S(t_m - s)B(X(t_m))dW(s).\end{aligned}$$

Recall that

$$\begin{aligned}&X_m^h \\ &= e^{\Delta t A_h}X_{m-1}^h + A_h^{-1}\left(e^{\Delta t A_h} - I\right)P_h F(X_{m-1}^h) + \int_{t_{m-1}}^{t_m} e^{(t_m - s)A_h}P_h B(X_{m-1}^h)dW(s) \\ &= e^{\Delta t A_h}X_{m-1}^h + \int_0^{\Delta t} e^{(\Delta t - s)A_h}P_h F(X_{m-1}^h)ds + \int_{t_{m-1}}^{t_m} e^{(t_m - s)A_h}P_h B(X_{m-1}^h)dW(s) \\ &= S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1}\left(\int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h F(X_k^h)ds + \int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h B(X_k^h)dW(s)\right) \\ &= S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1}\left(\int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h F(X_k^h)ds\right) \\ &\quad + \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h B(X_k^h)dW(s) \\ &= Z_m^h + O_m^h,\end{aligned}$$

with

$$Z_m^h = S_h(t_m)P_h X_0 + \sum_{k=0}^{m-1}\left(\int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h F(X_k^h)ds\right).$$

We examine the error

$$\begin{aligned}X(t_m) - X_m^h &= \overline{X}(t_m) + O(t_m) - X_m^h \\ &= \overline{X}(t_m) + O(t_m) - \left(Z_m^h + O_m^h\right) \\ &= \left(\overline{X}(t_m) - Z_m^h\right) + \left(O(t_m) - O_m^h\right) \\ &= I + II, \qquad (7.12)\end{aligned}$$

thus

$$\mathbf{E}\|X(t_m) - X_m^h\|^2 \leq 2\left(\mathbf{E}\|I\|^2 + \mathbf{E}\|II\|^2\right). \tag{7.13}$$

We follow the same approach as in Chapter 6. Let us estimate the first term $\mathbf{E}\|I\|^2$. Using the definition of T_h from (6.8), the first term I can be expanded

$$\begin{aligned} I &= T_h X_0 + \sum_{k=0}^{m-1} \int_{t_k t}^{t_{k+1}} S(t_m - s)F(X(s)) - S_h(t_m - s)P_h F(X_k^h)ds \\ &= T_h X_0 + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h(F(X(t_k)) - F(X_k^h)))ds \\ &\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_h(t_m - s)P_h(F(X(s)) - F(X(t_k)))ds \\ &\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (S(t_m - s) - S_h(t_m - s)P_h)F(X(s))ds \\ &= I_1 + I_2 + I_3 + I_4. \end{aligned} \tag{7.14}$$

Then

$$\mathbf{E}\|I\|^2 \leq 4\left(\mathbf{E}\|I_1\|^2 + \mathbf{E}\|I_2\|^2 + \mathbf{E}\|I_3\|^2 + \mathbf{E}\|I_4\|^2\right).$$

Let us estimate I_1, for $0 \leq \gamma < 1$ with $2\gamma \leq r$ and $r \in \{1, 2\}$, if $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, equation (6.9) of Lemma 6.4 with $\beta = 2\gamma$ yields

$$\mathbf{E}\|I_1\|^2 \leq C t_m^{-(r-2\gamma)} h^{2r}\left(\mathbf{E}\|X_0\|_{2\gamma}^2\right),$$

and if $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ we have

$$\mathbf{E}\|I_1\|^2 \leq C h^4\left(\mathbf{E}\|X_0\|_2^2\right).$$

For the term I_2, using Assumption 4.12, the triangle inequality as well as the fact that

$S_h(t)$ and P_h are bounded operators with Fubini's theorem yields

$$\begin{aligned}
\mathbf{E}\|I_2\|^2 &\leq Cm\sum_{k=0}^{m-1}\mathbf{E}\| \int_{t_k}^{t_{k+1}} S_h(t_m-s)P_h\left(F(X(t_k))-F(X_k^h)\right)ds\|^2 \\
&\leq Cm\sum_{k=0}^{m-1}\mathbf{E}\left(\int_{t_k}^{t_{k+1}} \|F(X(t_k))-F(X_k^h)\|ds\right)^2 \\
&\leq Cm\,\Delta t\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k)-X_k^h\|^2\right)ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k)-X_k^h\|^2\right)ds.
\end{aligned}$$

The Lipschitz condition, the triangle inequality, the fact that S_h and P_h are bounded, together with Lemma 7.3 yields

$$\begin{aligned}
&(\mathbf{E}\|I_3\|^2)^{1/2} \\
&\leq \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (\mathbf{E}\|S_h(t_m-s)P_h(F(X(s))-F(X(t_k))\|^2)^{1/2}ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (\mathbf{E}\|F(X(s))-F(X(t_k))\|)^{1/2}ds \\
&\leq C\left(\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (s-t_k)^{\gamma/2}ds\right) \\
&\quad\times\left(\mathbf{E}\|X_0\|_\gamma^2+\mathbf{E}\left(\sup_{0\leq s\leq T}(1+\|X(s)\|)\right)^2+\left(\sup_{0\leq s\leq T}\mathbf{E}\,(1+\|X(s)\|)^2\right)\right)^{1/2} \\
&\leq C\Delta t^{\gamma/2}\left(\mathbf{E}\|X_0\|_\gamma^2+\mathbf{E}\left(\sup_{0\leq s\leq T}(1+\|X(s)\|)\right)^2+\left(\sup_{0\leq s\leq T}\mathbf{E}\,(1+\|X(s)\|)^2\right)\right)^{1/2},
\end{aligned}$$

thus

$$\mathbf{E}\|I_3\|^2 \leq C\Delta t^{\gamma}.$$

If $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ we obviously have $\mathbf{E}\|I_3\|^2 \leq C\Delta t^{1-\epsilon}$ by taking $\gamma = 1-\epsilon$ in Lemma 7.3, $\epsilon \in (0, 1/2)$ small enough.

Notice that we have a similar estimation for the term I_3 in the case of additive noise (see Chapter 5 and Chapter 6).

Let us estimate $\left(\mathbf{E}\|I_4\|^2\right)^{1/2}$. For $r = 1, \beta = 0$ in Lemma 6.4 yields

$$\begin{aligned}\left(\mathbf{E}\|I_4\|^2\right)^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|T_h(t_m - s)F(X(s))\|^2\right)^{1/2} ds \\ &\leq Ch \sup_{0\leq s\leq T} \left(\mathbf{E}\|F(X(s))\|^2\right)^{1/2} \left(\int_0^{t_m} (t_m - s)^{-1/2}\right) \\ &\leq Ch,\end{aligned}$$

thus

$$\mathbf{E}\|I_4\|^2 \leq Ch^2.$$

For $r = 2, \beta \in (0, 1)$ small enough in equation (6.9) of Lemma 6.4, if for $F(X(t)) \in L_2\left(\mathbb{D}, \mathcal{D}((-A)^{\beta/2})\right)$ we have

$$\begin{aligned}\left(\mathbf{E}\|I_4\|^2\right)^{1/2} &\leq \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|T_h(t_m - s)F(X(s))\|_\beta^2\right)^{1/2} ds \\ &\leq Ch^2 \sup_{0\leq s\leq T} \left(\mathbf{E}\|F(X(s))\|_\beta^2\right)^{1/2} \left(\int_0^{t_m} (t_m - s)^{-1+\beta/2}\right) \\ &\leq Ch^2,\end{aligned}$$

thus

$$\mathbf{E}\|I_4\|^2 \leq Ch^4.$$

Combining the previous estimates yields: For $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma)),\ 1/2 \leq \gamma < 1$

$$\mathbf{E}\|I\|^2 \leq C\left(h^2 + \Delta t^\gamma + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2\right) ds\right).$$

For $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma)),\ 0 < \gamma \leq 1/2$

$$\mathbf{E}\|I\|^2 \leq C\left(t_m^{-1+2\gamma} h^2 + \Delta t^\gamma + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2\right) ds\right).$$

For $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$, $0 < \gamma < 1$ and $F(X(t)) \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\beta/2}))$, $\beta \in (0, 1/2)$ small enough

$$\mathbf{E}\|I\|^2 \leq C \left(t_m^{-2+2\gamma} h^4 + \Delta t^\gamma + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2 \right) ds \right)$$

and $X_0 \in L_2(\mathbb{D}, \mathcal{D}(-A))$ and $F(X(t)) \in L_2(\mathbb{D}, \mathcal{D}((-A)^{\beta/2}))$, $\beta \in (0, 1/2)$ small enough

$$\mathbf{E}\|I\|^2 \leq C \left(h^4 + \Delta t^{1-\epsilon} + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(\mathbf{E}\|X(t_k) - X_k^h\|^2 \right) ds \right),$$

with $\epsilon \in (0, 1/2)$ small enough.

Now we look at the noise term, let us estimate $\mathbf{E}\|II\|^2$, we follow the same approach as in [87]. Note that in the case of additive noise the estimation is straightforward and weak assumption on the noise improve the accuracy [91]. For multiplicative noise we have

$$\begin{aligned}
II &= \int_0^{t_m} S(t_m - s) B(X(s) dW(s) - \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_h(t_m - t_k) P_h B(X_k^h) dW(s) \\
&= \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_h(t_m - t_k) P_h \left(B(X(t_k)) - B(X_k^h) \right) dW(s) \\
&\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_h(t_m - t_k) P_h (B(X(s)) - B(X(t_k)) dW(s) \\
&\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (S(t_m - t_k) - S_h(t_m - t_k) P_h) B(X(s)) dW(s) \\
&\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \left(S(t_m - s) - S(t_m - t_k) \right) B(X(s)) dW(s) \\
&= II_1 + II_2 + II_3 + II_4. \qquad (7.15)
\end{aligned}$$

Then

$$\mathbf{E}\|II\|^2 \leq 4 \left(\mathbf{E}\|II_1\|^2 + \mathbf{E}\|II_2\|^2 + \mathbf{E}\|II_3\|^2 + \mathbf{E}\|II_4\|^2 \right).$$

Let us estimate each term. Using the Ito isometry, the boundedness of S_h and P_h with

Assumption 4.13 yields

$$\begin{aligned}
\mathbf{E}\|II_1\|^2 &= \mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_h(t_m-t_k)P_h\left(B(X(t_k))-B(X_k^h)\right)dW(s)\|^2 \\
&= \sum_{k=0}^{m-1}\mathbf{E}\|\int_{t_k}^{t_{k+1}} S_h(t_m-t_k)P_h\left(B(X(t_k))-B(X_k^h)\right)dW(s)\|^2 \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \mathbf{E}\|\left(B(X(t_k))-B(X_k^h)\right)\|_{L_2^0}^2 ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \mathbf{E}\|X(t_k)-X_k^h\|^2 ds.
\end{aligned}$$

For $X_0 \in L_2\left(\mathbb{D}, \mathcal{D}\left((-A)^\gamma\right)\right)$, using Lemma 7.3 yields

$$\begin{aligned}
\mathbf{E}\|II_2\|^2 &= \mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} S_h(t_m-t_k)P_h(B(X(s))-B(X(t_k)))dW(s)\|^2 \\
&= \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \mathbf{E}\|S_h(t_m-t_k)P_h(B(X(s))-B(X(t_k)))\|_{L_2^0}^2 ds \\
&\leq C\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} \mathbf{E}\|X(s)-X(t_k)\|^2 ds \\
&\leq C\left(\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}} (s-t_k)^\gamma ds\right) \\
&\quad \times\left(\mathbf{E}\|X_0\|_\gamma^2 + \mathbf{E}\left(\sup_{0\leq s\leq T}\left(1+\|X(s)\|\right)\right)^2 + \left(\sup_{0\leq s\leq T}\mathbf{E}\|\left(1+X(s)\|\right)^2\right)\right) \\
&\leq C\Delta t^\gamma.
\end{aligned}$$

For $X_0 \in L_2\left(\mathbb{D}, \mathcal{D}(-A)\right)$, taking $\gamma = 1-\epsilon$, with ϵ small enough yields

$$\mathbf{E}\|II_2\|^2 \leq C\Delta t^{1-\epsilon}.$$

Let us estimate $\mathbf{E}\|II_3\|^2$. By Ito's isometry and Lemma 6.4, we have

$$\begin{aligned}\mathbf{E}\|II_3\|^2 &= \mathbf{E}\|\sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}(S(t_m-t_k)-S_h(t_m-t_k)P_h)B(X(s))dW(s)\|^2\\ &= \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}\mathbf{E}\|(S(t_m-t_k)-S_h(t_m-t_k)P_h)B(X(s))\|_{L_2^0}^2ds\\ &= \sum_{k=0}^{m-1}\int_{t_k}^{t_{k+1}}\mathbf{E}\|T_h(t_m-t_k)B(X(s))\|_{L_2^0}^2ds.\end{aligned}$$

Indeed using Lemma 6.4 with $r=1,\ \beta=\alpha\in(0,1/2)$, if $t\in[0,T]$, $b(X(t))\in L_2\left(\mathbb{D},\mathcal{D}((-A)^{\alpha/2})\right)$, α small enough we have

$$\begin{aligned}\|T_h(t_m-t_k)B(X(s))\|_{L_2^0}^2 &= \sum_{i\in\mathbb{N}}^{\infty}\|T_h(t_m-t_k)b(X(s))Q^{1/2}e_i\|^2\\ &= \sum_{i\in\mathbb{N}}^{\infty}\|T_h(t_m-t_k)b(X(s))\|^2\|Q^{1/2}e_i\|^2\\ &\le Ch^2(t_m-t_k)^{-1+\alpha}\|b(X(s))\|_\alpha^2\mathbf{Tr}(Q),\end{aligned}$$

thus

$$\mathbf{E}\|II_3\|^2\le Ch^2\mathbf{Tr}(Q)\sup_{0\le s\le T}\mathbf{E}\|b(X(s))\|_\alpha^2\left(\Delta t^\alpha\sum_{k=0}^{m-1}(m-k)^{-1+\alpha}\right),$$

since

$$\Delta t^\alpha\sum_{k=0}^{m-1}(m-k)^{-1+\alpha},$$

is the discrete form of

$$\Delta t^\alpha\int_0^{m-1}(m-s)^{-1+\alpha}ds\le\Delta t^\alpha M^\alpha=T^\alpha,$$

we therefore have

$$\mathbf{E}\|II_3\|^2\le Ch^2\mathbf{Tr}(Q)\sup_{0\le s\le T}\mathbf{E}\|b(X(s))\|_\alpha^2.$$

For $b(X(t))\in L_2\left(\mathbb{D},\mathcal{D}(-A)\right)$ we obviously have using Lemma 6.4 with $r=2,\ \beta=0$

$$\mathbf{E}\|II_3\|^2\le Ch^4\mathbf{Tr}(Q)\sup_{0\le s\le T}\mathbf{E}\|b(X(s)\|_2^2.$$

Let us estimate $\mathbf{E}\|II_4\|^2$, for $b(X(t)) \in L_2\left(\mathbb{D}, \mathcal{D}((-A)^{\alpha/2})\right)$, $\alpha \in (0, 1/2)$ small enough, the following estimation holds

$$\begin{aligned}
\mathbf{E}\|II_4\|^2 &= \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \mathbf{E}\|(S(t_m - s) - S(t_m - t_k))B(X(s))\|_{L_2^0}^2 ds \\
&\leq \mathbf{Tr}(Q) \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \|(-A)^{-\alpha/2} S(t_m - s) - S(t_m - t_k))\|_{L(L^2(\Omega))}^2 ds \\
&\quad \times \left(\sup_{0 \leq s \leq T} \mathbf{E}\|b(X(s))\|_\alpha^2 \right),
\end{aligned}$$

since

$$\begin{aligned}
&\|(-A)^{-\alpha/2}(S(t_m - s) - S(t_m - t_k))\|_{L(L^2(\Omega))}^2 \\
&= \|(-A)^{(1-\alpha)/2} S(t_m - s)(-A)^{(-1/2)}(\mathbf{I} - S(s - t_k))\|_{L(L^2(\Omega))}^2 \\
&\leq C(s - t_k)(t_m - s)^{(\alpha - 1)},
\end{aligned}$$

thus

$$\begin{aligned}
\mathbf{E}\|II_4\|^2 &\leq C \left(\sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (s - t_k)(t_m - s)^{(\alpha-1)} ds \right) \left(\sup_{0 \leq s \leq T} \mathbf{E}\|b(X(s))\|_\alpha^2 \right) \\
&\leq \Delta t \left(\sup_{0 \leq s \leq T} \mathbf{E}\|b(X(s))\|_\alpha^2 \right).
\end{aligned}$$

Combining previous estimations related to II yields: For $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$ and $b(X(t)) \in L_2\left(\mathbb{D}, \mathcal{D}((-A)^{\alpha/2})\right)$, $\alpha > 0$ small enough,

$$\mathbf{E}\|II\|^2 \leq C \left(h^2 + \Delta t^\gamma + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \mathbf{E}\|X(t_k) - X_k^h\|^2 ds \right).$$

For $X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\gamma))$ and $b(X(t)) \in L_2(\mathbb{D}, \mathcal{D}(-A))$

$$\mathbf{E}\|II\|^2 \leq C \left(h^4 + \Delta t^\gamma + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} \mathbf{E}\|X(t_k) - X_k^h\|^2 ds \right).$$

Combining the estimates of $\mathbf{E}\|I\|^2$ and $\mathbf{E}\|II\|^2$ and applying the discrete Gronwall's lemma ends the proof. ■

7.3.3 Proof of Theorem 7.2 for the scheme SETDM0

We just give a sketch of the mains steps. Recall that

$$\begin{aligned}
Y_m^h &= e^{\Delta t A_h}\left(Y_{m-1}^h + \Delta t P_h F(Y_{m-1}^h)\right) + \int_{t_{m-1}}^{t_m} e^{\Delta t A_h} P_h B(Y_{m-1}^h) dW(s) \\
&= e^{\Delta t A_h} Y_{m-1}^h + \int_0^{\Delta t} e^{\Delta t\, A_h} P_h F(Y_{m-1}^h) ds + \int_{t_{m-1}}^{t_m} e^{\Delta A_h} P_h B(Y_{m-1}^h) dW(s) \\
&= S_h(t_m) P_h X_0 + \sum_{k=0}^{m-1} \left(\int_{t_k}^{t_{k+1}} S_h(t_m - t_k) P_h F(Y_k^h) ds \right. \\
&\quad \left. + \int_{t_k}^{t_{k+1}} S_h(t_m - t_k) P_h B(Y_k^h) dW(s) \right) \\
&= S_h(t_m) P_h X_0 + \sum_{k=0}^{m-1} \left(\int_{t_k}^{t_{k+1}} S_h(t_m - t_k) P_h F(Y_k^h) ds \right) \\
&\quad + \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} S_h(t_m - t_k) P_h B(Y_k^h) dW(s) \\
&= z_m^h + o_m^h.
\end{aligned}$$

We can therefore put the estimation of the error in form of (7.12), the estimate of the corresponding $\mathbf{E}\|I\|^2$ is the same as in Theorem 7.2 with the extra term

$$I_5 = \sum_{k=0}^{m-1} \int_{t_k}^{t_{k+1}} (S(t_m - s) - S(t_m - t_k)) F(X(s)) ds.$$

This is estimated in Theorem 5.7 of Chapter 5 as

$$\left(\mathbf{E}\|I_5\|^2\right)^{1/2} < C\left(\Delta t + \Delta t |\ln(\Delta t)|\right) \leq C\Delta t^{\gamma/2}.$$

The estimation of $\mathbf{E}\|II\|^2$ is the same as in Theorem 7.2 for SETDM1 scheme.

7.4 Simulations

Efficient implementation of $\varphi_i,\ i = 0, 1$ can be achieved by either the real fast Léja points technique or the Krylov subspace technique that we have presented in Chapter 3.

In Section 7.4.1 we apply the scheme to a linear problem where we can construct the exact solution for the truncated noise. The finite element method is used for space discretization. We use the real fast Léja point technique to compute the exponential functions $\varphi_i,\ i = 0, 1$. We use the noise with exponential correlation which is obviously a trace class noise.

In Section 7.4.2 we apply the scheme to nonlinear stochastic flow with multiplicative noise in heterogeneous media. To deal with high Péclet number flow, we use the finite volume method for space discretization. We use the Krylov subspace technique to compute the exponential functions $\varphi_i,\ i = 0, 1$, implemented in the matlab functions `expv.m` and `phiv.m` of the package Expokit [50]. Here we use the H^{r_1} noise as in the previous chapters.

In the legends of our graphs, "SETDM1" denotes results from the SETDM1 scheme, "SETDM0 " denotes results from the SETDM0 scheme and "Implicit" denotes results from the standard semi implicit Euler-Maruyama scheme.

7.4.1 Example 1

As a simple example consider the reaction diffusion equation in the time interval $[0, T]$ with diffusion coefficient $D > 0$

$$dX = (D\Delta X - 0.5X)dt + dW \qquad X(0) = X_0, \qquad \Omega = [0, L_1] \times [0, L_2]$$

with homogeneous Neumann boundary condition. We take f in the equation (7.3) to be linear here

$$f(u) = -0.5u. \tag{7.16}$$

The corresponding Nemytskii operator F is obtained from (7.5). Of course, in general, F will be nonlinear. Here $b(x, u) = 1,\ x \in \Omega,\ u \in \mathbb{R}$.

As in the previous chapters, we consider the covariance operator Q with the following covariance function (kernel) which is strongly exponential decay

$$C_r((x_1, y_1); (x_2, y_2)) = \frac{\Gamma}{4b_1b_2} \exp\left(-\frac{\pi}{4}\left[\frac{(x_2 - x_1)^2}{b_1^2} + \frac{(y_2 - y_1)^2}{b_2^2}\right]\right)$$

where b_1, b_2 are spatial correlation lengths in $x-$ axis and y- axis respectively and $\Gamma > 0$.

The eigenfunctions $\{e_i^{(1)}e_j^{(2)}\}_{i,j\geq 0}$ of the operator $A = D\Delta$ is given by

$$\begin{cases} e_0^{(l)} = \sqrt{\dfrac{1}{L_l}}, \quad \lambda_0^{(l)} = 0, \quad e_i^{(l)} = \sqrt{\dfrac{2}{L_l}}\cos(\lambda_i^{(l)}x), \quad \lambda_i^{(l)} = \dfrac{i\,\pi}{L_l} \\ \\ l \in \{1,2\} \quad i = 1,2,3,\cdots \end{cases} \tag{7.17}$$

with the corresponding eigenvalues $\{\lambda_{i,j}\}_{i,j\geq 0}$ given by

$$\lambda_{i,j} = (\lambda_i^{(1)})^2 + (\lambda_j^{(2)})^2.$$

Recall that the corresponding values of $\{q_{i,j}\}_{i+j>0}$ in the representation (4.3) are given by

$$q_{i,j} = \Gamma \exp\left[-\frac{1}{2\pi}\left((\lambda_i^{(1)}b_1)^2 + (\lambda_j^{(2)}b_2)^2\right)\right].$$

We compute the exponential functions $\varphi_i,\ i = 0, 1$ with the real fast Léja point technique and the absolute tolerance 10^{-6}. In our simulation we take $L_1 = L_2 = 1$ and the finite element mesh is contructed from the rectangular grid with size $\Delta x = \Delta y = 1/150$. Figure 7.1(a) shows the time convergence of SETDM1, SETDM0 and semi implict schemes. The three methods have the same order of accuracy. The temporal order of convergence is 0.9 for all the schemes, the order is high compared to the predicted order of 0.5 in Theorem 5.7. This is explained by the fact that F is linear and therefore belongs to the class of functions satisfying the condition (a) of Assumption 5.2 in Chapter 5, which allowed high order accuracy in the schemes built in Chapter 5 and Chapter 6 using linear functionals of the noise and by the fact that the noise is smooth.

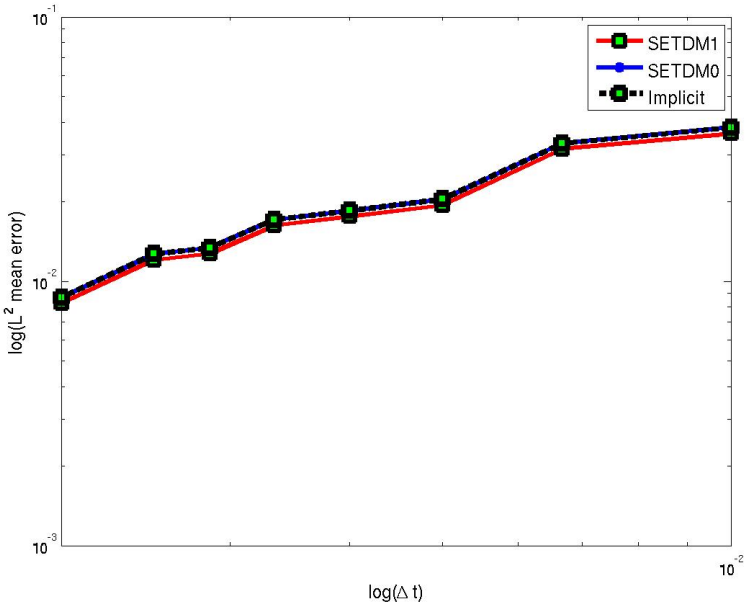
7.4.2 Example 2

As a more challenging example we consider the stochastic advection diffusion reaction SPDE

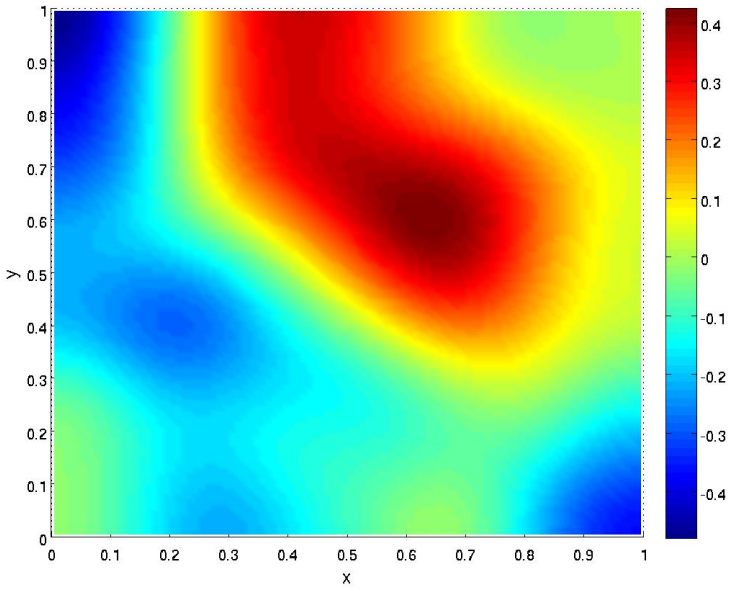
$$dX = \left(\nabla \cdot \mathbf{D}\nabla X - \nabla \cdot (\mathbf{q}X) - \frac{X}{|X|+1}\right)dt + XdW, \quad \Omega = [0,1]\times[0,1] \tag{7.18}$$

$$\mathbf{D} = \begin{pmatrix} 10^{-2} & 0 \\ 0 & 10^{-3} \end{pmatrix} \tag{7.19}$$

with mixed Neumann-Dirichlet boundary conditions and constant velocity $\mathbf{q} = (1, 0)$ for homogeneous medium. According to Theorem 7.8, we need to take the initial data



(a)



(b)

Figure 7.1: (a) Convergence of the root mean square L^2 norm at $T = 1$ as a function of Δt with 10 realizations with $X_0 = 0,\ \Gamma = 1,\ D = 1$. The noise is white in time and with exponential correlation in space with lengths $b_1 = b_2 = 0.2$. The observed temporal order of convergence in time is 0.9 for all schemes. In (b) we plot a sample of true solution.

$X_0 \in L_2(\mathbb{D}, \mathcal{D}((-A)^\beta)),\ \beta > 0$ to have a regular solution such that XdW make sense. For our simulation we take $X_0 = 0$. In terms of equation (7.3) the nonlinear terms f and b are given by

$$f(x,u) = -\frac{u}{(|u|+1)}, \qquad b(x,u) = u, \quad u \in \mathbb{R},\ x \in \Omega \tag{7.20}$$

and the corresponding Nemytskii operators F and B are obtained from (7.5) and clearly satisfy Assumption 4.12 (if the domain of f is restricted to $\mathbb{R}^+$) and Assumption 4.13 (see [95, Section 4]) respectively, where (7.17) is used in the noise representation (4.3).

The linear operator A is given by

$$A = \nabla \cdot \mathbf{D}\nabla(.) - \nabla \cdot \mathbf{q}(.). \tag{7.21}$$

For a heterogeneous medium we considered three parallel high permeability streaks as in Chapter 6. This could represent for example a highly idealized fracture pattern. We use here the noise in H^{r_1} as in Chapter 5, where we take the following values for $\{q_{i,j}\}_{i+j>0}$ in the representation (4.3)

$$q_{i,j} = \Gamma / \left(i+j\right)^{r_1/2}, \qquad r_1 > 0 \tag{7.22}$$

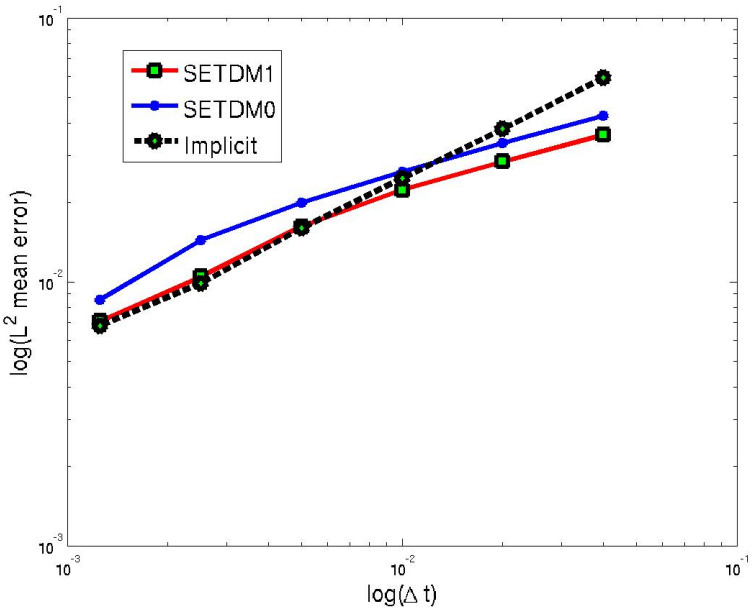
with (7.17). In our simulation we use $r_1 = 0.25,\ \Gamma = 0.02$. To deal with high Péclet flows we discretize in space using finite volumes. We can write the semi-discrete finite volume of (7.18) as

$$dX^h = (A_h X^h + P_h F(X^h)) + P_h B(X^h) dW. \tag{7.23}$$

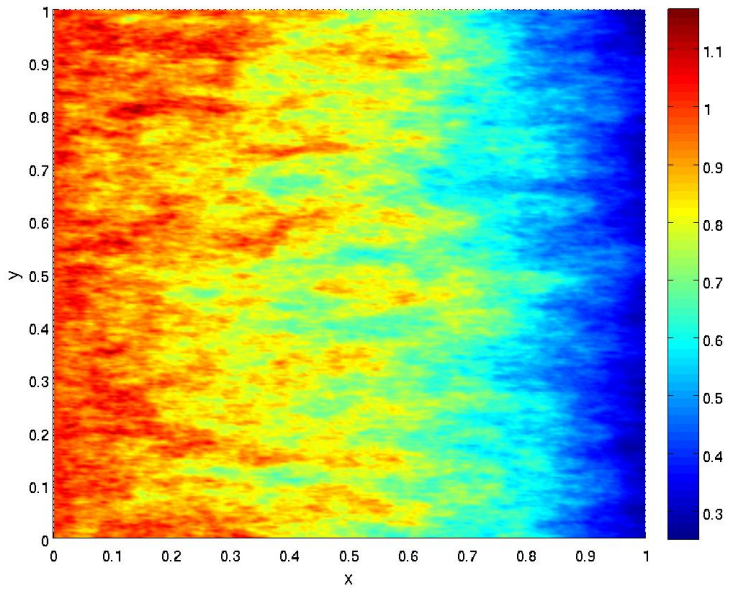
We compute the exponential matrix functions φ_i with the Krylov subspace technique with dimension $m = 6$ and the absolute tolerance 10^{-6}. We use a rectangular finite volume mesh and perform our simulation by following the sample paths as in [98].

Figure 7.2(a) shows the convergence of SETDM0, SETDM1 and semi implicit schemes for homogeneous porous medium. The scheme SETDM1 seems to be more accurate for large time steps but for small time steps it has the same order of accuarcy as the semi implicit scheme. The observed temporal order is 0.49 for SETDM1 scheme, 0.48 for SETDM0 scheme and 0.58 for the semi implicit scheme. We used only 30 realizations and the convergence order is close to the 0.5, the predicted order of convergence in Theorem 5.7. A sample of "true solution" is shown in Figure 7.2(b) with $\Delta t = 1/1600$.

Figure 7.3(a) shows the convergence of SETDM0 and SETDM1 schemes for heterogeneous porous medium. It also shows that SETDM1 is more accurate than SETDM0 scheme. The observed temporal order is 0.51 for SETDM1 scheme and 0.55 for SETDM0 scheme. We only used 30 realizations, more realizations will probably give the convergence order close to 0.5, the predicted order in Theorem 5.7. A sample of a "true solution" is shown in Figure 7.3(b) with $\Delta t = 1/1600$ while the mean of the "true solution" for 30 realizations is shown in Figure 7.4(b). Figure 7.4(a) shows the streamline of the velocity field.

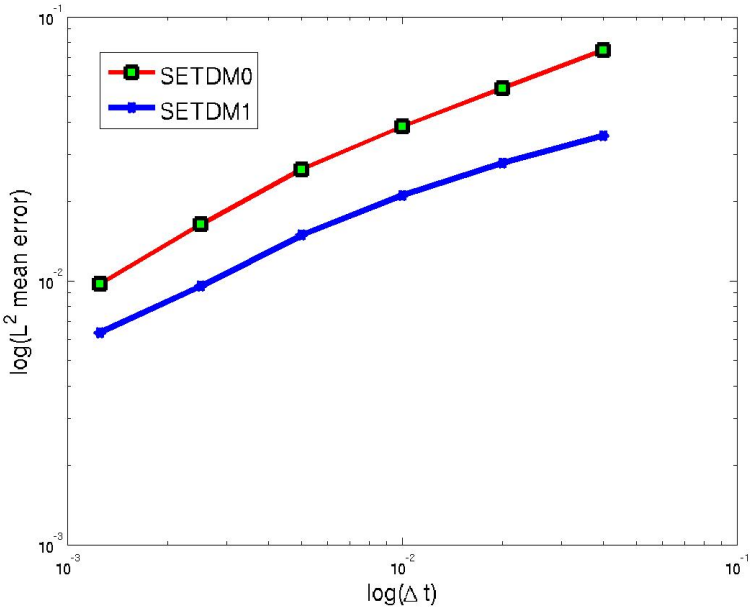


(a)

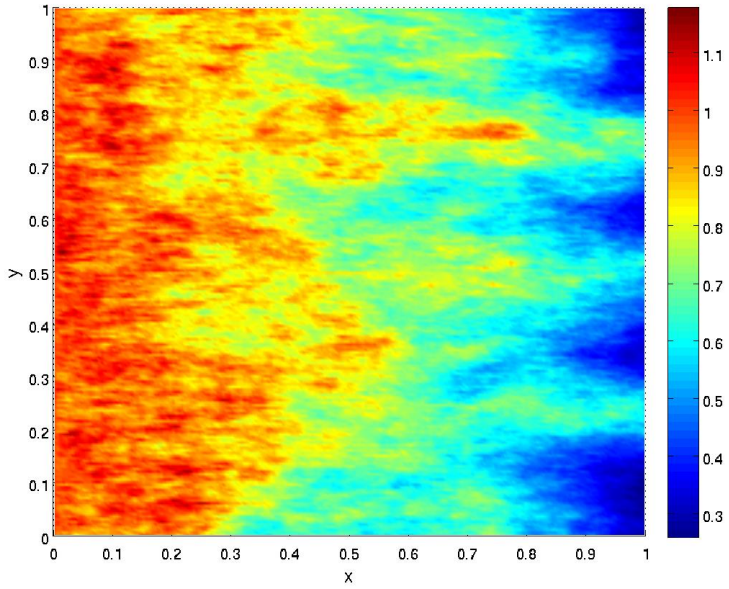


(b)

Figure 7.2: (a) Convergence of the root mean square L^2 norm at $T = 1$ as a function of Δt with 30 realizations with $\Delta x = \Delta y = 1/300$, $X_0 = 0$, $\Gamma = 0.02$ for homogeneous medium. The noise is white in time and in H^{r_1} in space, $r_1 = 0.25$. The temporal order of convergence in time is $0.49, 0.48$ and 0.58 for SETD1, SETD0 and semi implicit schemes respectively. In (b) we plot a sample of a "true solution" for $r_1 = 0.25$ with $\Delta t = 1/1600$.

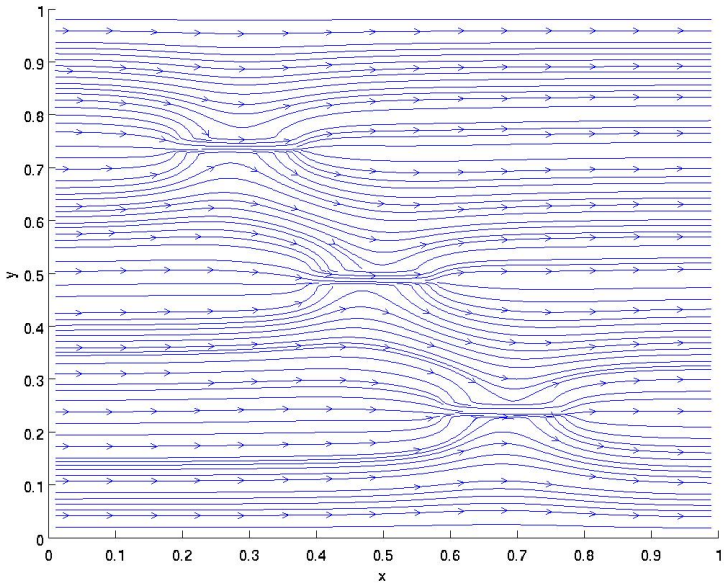


(a)

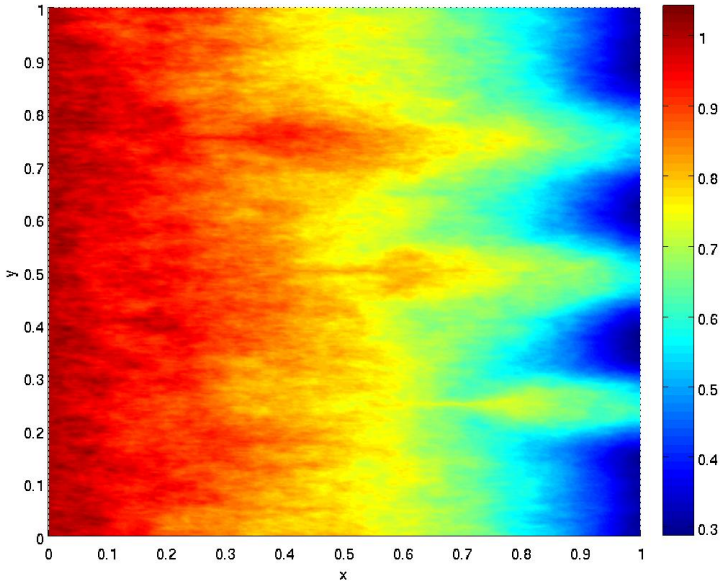


(b)

Figure 7.3: (a) Convergence of the root mean square L^2 norm at $T = 1$ as a function of Δt with 30 realizations with $\Delta x = \Delta y = 1/350$, $X_0 = 0$, $\Gamma = 0.02$ for heterogeneous medium. The noise is white in time and in H^{r_1} in space, $r_1 = 0.25$. The temporal orders of convergence in time are 0.51 for SETDM1 scheme and 0.55 for SETDM0. In (b) we plot a sample "true solution" for $r_1 = 0.25$ with $\Delta t = 1/1600$.



(a)



(b)

Figure 7.4: Streamline of the velocity and the mean of the "true solution" for 30 realizations corresponding to Figure 7.3. In (a) we plot the streamline of the velocity field while (b) shows the mean of the "true solution" for 30 realizations.

Conclusion

In this chapter we review the main contributions of this thesis. Our goal in this thesis was to develop efficient numerical schemes for deterministic and stochastic flow and transport in porous media

From Chapter 1 to Chapter 3, we considered two deterministic exponential integrator schemes: ETD1 and EEM. We gave the time and space convergence proof for ETD1 using the finite volume method for space discretization. Using the real fast Léja points and the Krylov subspace techniques to compute the exponential matrix functions, we illustrate with two and three dimensional simulations that the exponential integrators are generally more efficient and accurate for advection dominated deterministic flow and transport in heterogeneous anisotropic porous media compared to standard implicit and semi implicit schemes.

From Chapter 4 to Chapter 7, we developed new efficient stochastic schemes which are mainly the extension of the ETD1 and exponential Lawson schemes in the general framework of abstract nonlinear parabolic SPDEs. The finite element method is used for space discretization, although any non diagonal method such as finite volume or finite difference can be also used. In Chapter 5 and Chapter 6, we considered SPDEs with space time additive noise. We used the self adjoint part of the linear operator (diffusion part in the case of advection–diffusion–reaction equation) coupled with the linear functional of noise to compute accurately the variance of the noise and to create the new schemes, the so called modified semi–implicit Euler-Maruyama scheme, and two stochastic exponential integrators which, by theoretical proofs and simulations outperform the standard semi–implicit Euler-Maruyama scheme. Lastly, in Chapter 7 we considered the general SPDEs with the general noise (additive or multiplicative noise) and extended the deterministic ETD1 and exponential Lawson schemes to stochastic exponential integrators using standard

Brownian increments for the noise. In all our schemes, convergence proofs have been given and for stochastic exponential integrators, the real fast Léja points and the Krylov subspace techniques have been used to compute the matrix exponential functions of the non diagonal matrix. The schemes have been applied to two dimensional stochastic flow and transport.

Our future direction will be to develop new schemes with high order accuracy for SPDEs with multiplicative noise. Also we will build software for flow and transport with exponential integrators using the real fast Léja points technique.

Bibliography

[1] M. A. Christie and M. J. Blunt. Tenth SPE comparative solution project: A Comparison of upscaling techniques. *SPE Reservoir Evaluation and Engineering*, 4(4),308-317, 2001.

[2] S. K. Matthäi and M. Belayneh. Fluid flow partitioning between fractures and a permeable rock matrix. *Geophys. Res. Lett.*, 31(7) L07602 doi:10.1029/2003GL019027, 2004.

[3] B. Berkowitz, A. Cortis, M. Dentz, and H. Scher. Modeling non-Fickian transport in geological formations as a continuous time random walk. *Geophys. Res. Lett.*, 44(2) RG2003 doi:10.1029/2005RG000178, 2006.

[4] O. Cirpka and P. Kitanidis. Characterization of mixing and dilution in heterogeneous aquifers by means of local temporal moments. *Water Resour. Res.*, 36(5) 1221–1236, 2000.

[5] A. M. Tartakovsky, G. Redden, P. L. Lichtner, T. D. Scheibe, and P. Meakin. Mixing-induced precipitation: experimental study and multi-scale numerical analysis. *Water Resour. Res.*, 44(6) W06S04 doi:10.1029/2006WR005725, 2008.

[6] A. M. Tartakovsky, G. D. Tartakovsky, and T. D. Scheibe. Effects of incomplete mixing on multicomponent reactive transport. *Adv. Water Resour.*, 31(11) 1674–1679, 2009.

[7] M. Christie, V. Demyanov, and D. Ebras. Uncertainty quantification for porous media flows. *J. Comput. Phys.*, 217(1) 143–158, 2006.

[8] B. Minchev and W. Wright. A review of exponential integrators for first order semi-linear problems. 2005.

[9] A. K. Kassam and L. N. Trefethen. Fourth-order time stepping for stiff PDES. *SIAM J. Comput.*, 26(4) 1214–1233, 2005.

[10] C. Moler and C. Van Loan. Ninteen Dubious Ways to compute the Exponential of a Matrix, twenty–five years later. *SIAM Review*, 45(1), pp. 3–49, 2003.

[11] P. Knabner and L. Angermann. Numerical methods for elliptic and parabolic partial differential equations solution. Springer, 2000.

[12] D. Henry. *Geometric theory of semilinear parabolic equations*. Number 840 in Lecture notes in mathematics. Springer, 1981.

[13] M. Caliari, M. Vianello, and L. Bergamaschi. The LEM exponential integrator for advection–diffusion–reaction equations. *J. Comput. Appl. Math.*, 210(1-2) 56–63, 2007.

[14] S. M. Cox and P. C. Matthews. Exponential time differencing for stiff systems. *J. Comput. Phys.*, 176(2) 430–455, 2002.

[15] A. Tambue, G. J. Lord, and S. Geiger. An exponential integrator for advection-dominated reactive transport in heterogeneous porous media. *Journal of Computational Physics*, 229(10):3957 – 3969, 2010.

[16] A. Tambue, S. Geiger, and G. J. Lord. Exponential time integrators for 3D Reservoir Simulation. In proceedings of the 12th European Conference on the Mathematics of Oil Recovery, Oxford, UK, 2010.

[17] H. Fujita and T. Suzuki. Evolutions problems (part1). in: P. G. Ciarlet, J. L. Lions (Eds.) Handbook of Numerical Analysis, vol II, North-Holland, Amsterdam, pp. 789–928, 1991.

[18] G. J. Lord and A. Tambue. A modified semi–implict Euler-Maruyama scheme for finite element discretization of SPDEs. arXiv:1004.1998v1, 2010.

[19] G. J. Lord and A. Tambue. Stochastic exponential integrators for finite element discretization of SPDEs with additive noise. http://arxiv.org/abs/1005.5315, 2010.

[20] P. B. Bedient, H. S. Rifai, and C. J. Newell. Ground Water Contamination: Transport and Remediation. Prentice Hall PTR , Englewood Cliffs, New Jersey 07632, 1994.

[21] Jacob Bear. Dynamics of fluids in porous media,. PT.1 and PT.2, American elsevier, 1983.

[22] M. K. Hubbert. Darcy's law and field equations of the flow of underground fluids. *Transactions of the American Institute of Mining,Metallurgical and Petroleum Engineers*, 207,222-239, 1956.

[23] L. Bucciarell, H. Einstein, H. Nepf, and S. Rudolph. 1.101 Introduction to Civil and Environmental Engineering Design I. Civil and Environmental Engineering, MIT, http://www.flickr.com/photos/mitopencourseware/sets/72157614684644687/, Fall 2006.

[24] G. Chavent, B. Cockburn, G. Cohen, and J. Jaffré. Une méthode d' eléments finis pour la simulaton dans un réservoir de déplacements bidimensionnel d'huile par de l'eau. rapport INRIA, No 353, 1985.

[25] O. Banton and L. Bangoy. Hydrogéologie. *Multiscience environnementale des eaux souterraines*, Presse de l'Universite de Québec, 1997.

[26] G. Da Prato and J. Zabczyk. *Stochastic Equations in Infinite Dimensions*, volume 44 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, Cambridge, 1992.

[27] R. G. Ghanem and P. D. Spanos. Stochastic finite elements. Springer Verlag, New York, 2003.

[28] F. Brezzi and M. Fortin. Mixed and Hybrid Finite Element Methods. Springer Verlag, New York, 1991.

[29] O. G. Ernst, C. E. Powell, D. J. Silvester, and E. Ullmann. Efficient Solvers for a Linear Stochastic Galerkin Mixed Formulation of Diffusion Problems with Random Data. *SIAM Journal Sci. Comp*, 31(2):1424–1447, 2009.

[30] Wuan Luo. *Wiener Chaos Expansion and Numerical Solutions of Stochastic Partial Differential Equations*. PhD thesis, California Institute of Technology, 2006.

[31] V. Thomée. Galerkin finite element methods for parabolic problems. Springer Series in Computational Mathematics, 1997.

[32] S. Larsson. Nonsmooth data error estimates with applications to the study of the long-time behavior of finite element solutions of semilinear parabolic problems,. Preprint 1992-36, Department of Mathematics, Chalmers University of Technology, 1992.

[33] A. Pazy. *Semigroups of linear operators and applications to partial differential equations*, volume 44 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 1983.

[34] R. Eymard, T. Gallouet, and R. Herbin. A cell-centred finite-volume approximation for anisotropic diffusion operators on unstructured meshes in any space dimension. *MA Journal of Numerical Analysis*, 26, 326–353, 2006.

[35] R. Eymard, T. Gallouet, and R. Herbin. Finite volume methods. in: P.G.Ciarlet, J.L.Lions (Eds.)Handbook of Numerical Analysis Volume 7, North-Holland, Amsterdam, pp. 713–1020, 2003.

[36] G. L. G. Sleijpen and D. R. Fokkema. Bicgstab(l) for linear equations involving unsymmetric matices with complex spectrum. *Electronic Transactions on Numerical Analysis.*, Volume 1, 11–32, 1993.

[37] K. W. Morton. Numerical solution of convection-diffusion problems. Chapman and Hall, 1996.

[38] A. Tambue. Temporal and spatial finite volume convergence of an Exponential integrator for local Lipschitz nonlinear reaction transport and applications to porous media. In preparation.

[39] D. A. Knoll, L. Chacon, L. G. Margolin, and V. A. Mousseau. On balanced approximations for time integration of multiple time scale systems. *J. Comput. Phys.*, 185(2) 583–611, 2003.

[40] D. L. Ropp, J. N. Shadid, and C. C. Ober. Studies of the accuracy of time integration methods for reaction–diffusion equations. *J. Comput. Phys.*, 94(2) 544–577, 2004.

[41] U. M. Ascher, S. J. Ruuth, and B. T. R. Wetton. Implicit explicit methods for time dependent partial differential equations. *SIAM J. Numer. Anal.*, 32(3) 797–823, 1995.

[42] M. G. Gerritsen and L. J. Durlofsky. Modeling fluid flow in oil reservoirs. *Annu. Rev. Fluid Mech.*, 37 211–238, 2005.

[43] G. Strang. On the construction and comparison of difference schemes. *SIAM J. Numer. Anal.*, 5(3) 506–517, 1968.

[44] C. I. Steefel and K. T. B. MacQuarrie. Approaches to modeling of reactive transport in porous media. in: P. C. Lichtner, C. I. Steefel, and E. H. Oelkers (Eds.) Reviews in Mineralogy and Geochemistry Volume 34, Mineralogical Society of America, Chantilly, pp. 83–129, 1996.

[45] M. R. Thiele, R. P. Batycky, and F. Orr. Simulating flow in heterogeneous systems using streamtubes and streamlines. *SPE Reservoir Eng.*, 11(1) 5–12, 1996.

[46] G. DiDonato and M. J. Blunt. Streamline-based dual-porosity simulation of reactive transport and flow in fractured reservoirs. *Water Resour. Res.*, 40(4) W04203 doi:10.1029/2003WR002772, 2004.

[47] R. D. Hornung and J. A. Trangenstein. Adaptive mesh refinement and multilevel iteration for flow in porous media. *J. Comput. Phys.*, 136(2) 522–545, 1997.

[48] H. Karimabadi, J. Driscoll, Y. A. Omelchenko, and N. Omidi. A new asynchronous methodology for modeling of physical systems: breaking the curse of courant condition. *J. Comput. Phys.*, 205(2) 755–775, 2005.

[49] Y. A. Omelchenko and H. Karimabadi. Self-adaptive time integration of flux-conservative equations with sources. *J. Comput. Phys.*, 216(1) 179–194, 2006.

[50] R. B. Sidje. Expokit: A software package for computing matrix exponentials. *ACM Trans. Math. Software*, 24(1), pp.130–156, 1998.

[51] H. Berland, B. Skaflestad, and W. Wright. A matlab package for exponential integrators. *ACM Trans. Math. Software*, 33(1), Article No. 4, 2007.

[52] J. Baglama, D. Calvetti, and L. Reichel. Fast léja points. *Electron. Trans. Num. Anal.*, 7, 124–140, 1998.

[53] L. Bergamaschi, M. Caliari, and M. Vianello. The RELPM exponential integrator for FE discretizations of advection-diffusion equations. in: M. Bubak, G. D. Van Albada, P. Sloot (Eds.), Lecture Notes in Computer Sciences Volume 3039, Springer Verlag, Berlin Heidelberg, pp. 434-442, 2004.

[54] M. Hochbruck and C. Lubich. On Krylov subspace approximations to the matrix exponential operator. *SIAM J. Numer. Anal.*, 34(5), pp. 1911–1925., 1997.

[55] A. Martinez, L. Bergamaschi, M. Caliari, and M. Vianello. A massively parallel exponential integrator for advection-diffusion models. *J. Comput. Appl. Math.*, 231(1), pp. 82–91, 2009.

[56] L. Bergamaschi, M. Caliari, A. Martinez, and M. Vianello. Comparing Léja and Krylov approximations of large scale matrix exponentials. *Comput. Sci. – ICCS*, 3994, pp.685–692, 2006.

[57] M. Caliari, M. Vianello, and L. Bergamaschi. Interpolating discrete advection diffusion propagators at Léja sequences. *J. Comput. Appl. Math.*, 172(1), pp. 79–99, 2004.

[58] M. Caliari and A. Ostermann. Implementation of exponential Rosenbrock-type integrators. *Appl. Numer. Math.*, 59(3-4) 568–581, 2009.

[59] C. Gonzáleza, A. Ostermann, and M. Thalhammer. A second-order magnus-type integrator for nonautonomous parabolic problems. *J. Comput. Appl. Math.*, 189 142 – 156, 2006.

[60] M. Hochbruck, A. Ostermann, and J. Schweitzer. Exponential rosenbrock-type method. *SIAM J. Numer. Anal.*, 47(1) 786–803, 2008.

[61] J. A. Oguntuase. On an inequality of Gronwall. *Journal of Inequalities in Pure and Applied Mathematics*, 2(1) article 9, 2001.

[62] M. Hochbruck and A. Ostermann. Exponential Runge-Kutta methods for parabolic problems. *Appl. Numer. Math.*, 53 323–339, 2005.

[63] J. W. Thomas. Numerical partial differential equations: finite difference methods. Springer Verlag, Berlin Heidelberg New York, 1995.

[64] M. Caliari. Accurate evaluation of divided differences for polynomial interpolation of exponential propagators. *Computing*, 80(2) 189–201, 2007.

[65] A. McCurdy, K. C. Ng, and B. N. Parlett. Accurate computation of divided differences of the exponential function. *Math. Comp.*, 43(186), pp. 501–528, 1984.

[66] L. Reichel. Newton interpolation at Léja points. *BIT*, 30(2), pp.332–346, 1990.

[67] G. H. Golub and C. F. Van Loan. Matrix computations, third ed. *Johns Hopkins University Press*, Baltimore, 1996.

[68] C. Zoppou and J. H. Knight. Analytical solution of a spatially variable coefficient advection-diffusion equation in up to three dimensions. *Appl. Math. Model.*, 23(9) 667–685, 1999.

[69] K. Stüben. A review of algebraic multigrid. *J. Comput. Appl. Math.*, 128(1-2) 281–309, 2001.

[70] T. Hou and X. Wu. A multiscale finite element method for elliptic problems in composite materials and porous media. *J. Comput. Phys.*, 134(1) 169–189, 1997.

[71] Y. Efendiev, T. Hou, and X. Wu. Convergence of a nonconformal multiscale finite element method. *SIAM J. Numer. Anal.*, 37(3) 888–910, 2000.

[72] P. Jenny and H. A. Tchelepi. Multi-scale finite-volume method for elliptic problems in subsurface flow simulation. *J. Comput. Phys.*, 187(1) 47–67, 203.

[73] Claudia Prévôt and Michael Röckner. *A Concise Course on Stochastic Partial Differential Equations*. Springer, 2007. ISBN-10: 3540707808.

[74] P-L. Chow. *Stochastic Partial Differential Equations*. Applied Mathematics and nonlinear Science. Chapman & Hall / CRC, 2007. ISBN-1-58488-443-6.

[75] T. Shardlow. Numerical simulation of stochastic PDEs for excitable media. *J. Comput. Appl. Math.*, 175(2):429–446, March 2005.

[76] E. Hausenblas. Approximation for semilinear stochastic evolution equations. *Potential Analysis*, 18(2):141–186, 2003.

[77] E. J. Allen, S. J. Novosel, and Z. Zhang. Finite element and difference approximation of some linear stochastic partial differential equations. *Stochastics Rep.*, 64(1-2):117–142, 1998.

[78] M. Kovács, S. Larsson, and F. Lindgren. Strong convergence of the finite element method with truncated noise for semilinear parabolic stochastic equations with additive noise. *Numer. Algor.*, (53) 309–320, 2010.

[79] Y. Yan. Semidiscrete Galerkin Approximation for a Linear Stochastic Parabolic Partial Differential Equation Driven by an Additive Noise. *BIT Numerical Mathematics*, 44(4):829–847, December 2004. DOI:10.1007/s10543-004-3755-5.

[80] T. Shardlow. Numerical methods for stochastic parabolic PDEs. *Numer. Funct. Anal. Optim.*, 20(1-2):121–145, 1999.

[81] G. J. Lord and J. Rougemont. A numerical scheme for stochastic PDEs with Gevrey regularity. *IMA J. Numer. Anal.*, 24(4):587–604, 2004.

[82] A. Jentzen and P. E. Kloeden. Overcoming the order barrier in the numerical approximation of SPDEs with additive space-time noise. *Proc. R. Soc. A*, 465(2102):649–667, 2009.

[83] A. Jentzen. High order pathwise numerical approximations of SPDES with additive noise. unpublished manuscript, 2009.

[84] A. Jentzen. Pathwise Numerical Approximations of SPDEs. *Potential Analysis*, 31(4):375–404, 2009.

[85] M. A. Katsoulakis, G. T. Kossioris, and O. Lakkis. Noise regularization and computations for the 1-dimensional stochastic Allen–Cahn problem. *Interfaces and Free Boundaries*, 2007.

[86] G. T. Kossioris and G. E. Zouraris. Fully-discrete finite element approximations for a fourth-order linear stochastic parabolic equation with additive space-time white noise. *ESAIM*, 2010.

[87] Y. Yan. Galerkin finite element methods for stochastic parabolic partial differential equations. *SIAM J. Num. Anal.*, 43(4):1363–1384, 2005.

[88] C. M. Elliott and S. Larsson. Error estimates with smooth and nonsmooth data for a finite element method for the Cahn-Hilliard equation. *Math. Comp.*, 1992.

[89] A. Jentzen, P. E. Kloeden, and G. Winkel. Efficient Simulation of Nonlinear parabolic SPDEs with additive noise. *Annals of Applied Probability*, In review, 2009.

[90] G. J. Lord and T. Shardlow. Postprocessing for stochastic parabolic partial differential equations. *SIAM J. Numer. Anal.*, 45(2):870–889 (electronic), 2007.

[91] P. Kloeden, G. J. Lord, A. Neuenkirch, and T. Shardlow. The exponential integrator scheme for stochastic partial differential equations: Pathwise error bounds. Accepted to J. Comp. A. Math.

[92] García-Ojalvo, Jordi, Sancho, and M. José. *Noise in spatially extended systems*. Institute for Nonlinear Science. Springer-Verlag, New York, 1999.

[93] I. Gradstein and I. Ryzhik. Tables of integrals series and products. Academic Press New York, 1980.

[94] G. J. Lord and A. Tambue. Stochastic exponential integrators for finite element discretization of SPDEs for additive to multiplicative noises. In preparation.

[95] A. Jentzen and M. Röckner. Regularity analysis for stochastic partial differential equations with nonlinear multiplicative trace class noise. http://arxiv.org/abs/1005.4095v1, 2010, 2010.

[96] G. Da Prato and J. Zabczyk. Second order partial differential equations in hilbert spaces. London Mathematical Society, Lecture Notes, 293, Cambridge University Press, 2002.

[97] I. B. Adamu. *Numerical simulations of stochastic differential equations & the stochastic Swift–Hohenberg equation.* PhD thesis, Department of Mathematics, Heriot–Watt University, In preparation.

[98] D. J. Higham. An Algorithmic Introduction to Numerical Simulation of Stochastic Differential Equations. *SIAM REVIEW*, 43(3) 525–546, 2001.